

Fuentes-Albero, Cristina; Melosi, Leonardo

Working Paper

Methods for computing marginal data densities from the gibbs output

Working Paper, No. 2011-31

Provided in Cooperation with:

Department of Economics, Rutgers University

Suggested Citation: Fuentes-Albero, Cristina; Melosi, Leonardo (2011) : Methods for computing marginal data densities from the gibbs output, Working Paper, No. 2011-31, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:

<https://hdl.handle.net/10419/59469>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Methods for Computing Marginal Data Densities from the Gibbs Output*

Cristina Fuentes-Albero
Rutgers University

Leonardo Melosi
London Business School

July 2011

Abstract

We introduce two new methods for estimating the Marginal Data Density (MDD) from the Gibbs output, which are based on exploiting the analytical tractability condition. Such a condition requires that some parameter blocks can be analytically integrated out from the conditional posterior densities. Our estimators are applicable to densely parameterized time series models such as VARs or DFMs. An empirical application to six-variate VAR models shows that the bias of a fully computational estimator is sufficiently large to distort the implied model rankings. One estimator is fast enough to make multiple computations of MDDs in densely parameterized models feasible.

Keywords: Marginal likelihood, Gibbs Sampler, time series econometrics, Bayesian econometrics, reciprocal importance sampling.

JEL Classification: C11, C15, C16, C32

*Correspondence: Cristina Fuentes-Albero: Department of Economics, 75 Hamilton Street, Rutgers University, New Brunswick, NJ 08901: cfuentes@economics.rutgers.edu. Leonardo Melosi: London Business School, Regent's Park, Sussex Place, London NW1 4SA, United Kingdom: lmelosi@london.edu We thank Jesús Fernández-Villaverde, Frank Schorfheide, Lucrezia Reichlin, Herman van Dijk, Paolo Surico, and Daniel Waggoner for helpful comments

1 Introduction

Modern macroeconometric methods are based on densely parameterized models such as vector autoregressive models (VAR) or dynamic factor models (DFM). Densely parameterized models deliver a better in-sample fit. It is well-known, however, that such models can deliver erratic predictions and poor out-of-sample forecasts due to parameter uncertainty. To address this issue, Sims (1980) suggested to use priors to constrain parameter estimates by "shrinking" them toward a specific point in the parameter space. Provided that the direction of shrinkage is chosen accurately, it has been shown that densely parameterized models are extremely successful in forecasting. This explains the popularity of largely parameterized models in the literature (Stock and Watson, 2002, Forni, Hallin, Lippi, and Reichlin, 2003, Koop and Porter 2004, Korobilis, forthcoming, Banbura, Giannone, and Reichlin, 2010 and Koop, 2011).

The direction of shrinkage is often determined by maximizing the marginal likelihood of the data (see Carriero, Kapetanios and Marcellino, 2010 and Giannone et al., 2010), also called marginal data density (MDD). The marginal data density is defined as the integral of the likelihood function with respect to the prior density of the parameters. In few cases, the MDD has an analytical representation. When an analytical solution for this density is not available, we need to rely on computational methods, such as the Chib's method (Chib, 1995), estimators based on Reciprocal Importance Sampling principle (Gelfand and Dey, 1994), or the Bridge Sampling estimator (Meng and Wong, 2006). Since all these methods rely on computational methods to integrate the model parameters out of the posterior density, their accuracy quickly deteriorates as the dimensionality of the parameter space grows large. Hence, there is a tension between the need for using broadly parameterized models for forecasting and the accuracy in estimating the MDD which influences the direction of shrinkage.

This paper aims at mitigating this tension by introducing two MDD estimators (henceforth, Method 1 and Method 2) that exploit the information about models' analytical structure. While Method 1 follows the approach proposed by Chib (1995), Method 2 is based upon the Reciprocal Importance Sampling principle. Conversely to fully computational methods, Method 1 and Method 2 rely on analytical integration of some parameter blocks¹.

We provide a guide on how to apply the estimators to a wide range of time series models, such as Vector AutoRegressive Models (VARs), Reduced Rank Regression Models

¹Fiorentini, Planas, and Rossi (2011) show how to integrate scale parameters out of the likelihood using Kalman filtering and Gaussian quadrature for dynamic mixture models.

such as Vector Equilibrium Correction Models (VECMs), Markov-Switching VAR models (MS VARs), Time-Varying Parameter VAR models (TVP VARs), Dynamic Factor Models (DFMs), and Factor Augmented VAR models (FAVARs). We show that all these models satisfy the two conditions that are needed for applying our estimators. The first condition (henceforth, *sampling condition*) requires that the posterior density can be approximated via the Gibbs sampler. The second condition (henceforth, *analytical tractability condition*) states that there exists an integer $i \geq 2$ such that it is possible to analytically integrate out $(i - 1)$ parameter blocks $\{\theta_1, \dots, \theta_{i-1}\}$ from the conditional posterior densities $\theta_i | (\Theta_{-i}, Y, D)$ for $i \in \{1, \dots, m\}$, where $\Theta_{-i} \equiv \{\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_m\}$, Y is the sample data, and D is a set of unobservable model variables.

By means of a Monte Carlo experiment, we show that exploiting the *analytical tractability condition* leads to sizeable gains in accuracy and computational burden which quickly grow with the dimensionality of the parameter space of the model. We consider VAR(p) models, in the form studied by Villani (2009) and Del Negro and Schorfheide (2010) (i.e., the so-called mean-adjusted VAR models), from one up to four lags, $p = 1, \dots, 4$. We fit these four VAR models, under a single-unit-root prior², to six data sets where the number of observable variables ranges from one to six. It is compelling to focus on mean-adjusted VAR models because the true conditional predictive density $Y | (\theta_{i+1}, \dots, \theta_m)$, with $i \geq 2$,³ can be analytically derived in closed form⁴. We can compare the performance of our estimators with that of the estimator proposed by Chib (1995). In particular, for mean-adjusted VAR models, Method 1 and Chib's method only differ in the computation of the conditional predictive density. While Method 1 evaluates the analytical expression for the conditional predictive density, Chib's method approximates this density computationally via Monte Carlo integration. Therefore, we can quantify the accuracy gains associated with exploiting the *analytical tractability condition* by comparing the conditional predictive density estimated by Chib's method with its true value. This assessment would have not been possible, if we had used

²For a thorough description of such a prior, see Del Negro and Schorfheide (2010), section 2.2.

³Note that the conditional predictive density is a component of the MDD, $p(Y)$. One can see this by decomposing the MDD as follows:

$$\begin{aligned} p(Y) &= \int p(Y|\theta_1, \dots, \theta_m) p(\theta_1, \dots, \theta_m) d\theta_1 \dots d\theta_m \\ &= \int \left(\int p(Y|\theta_1, \dots, \theta_m) p(\theta_1, \dots, \theta_i | \theta_{i+1}, \dots, \theta_m) d\theta_1 \dots d\theta_i \right) p(\theta_{i+1}, \dots, \theta_m) d\theta_{i+1} \dots d\theta_m \end{aligned}$$

where $p(Y|\theta_1, \dots, \theta_m)$ is the likelihood function and $p(\theta_1, \dots, \theta_i | \theta_{i+1}, \dots, \theta_m) p(\theta_{i+1}, \dots, \theta_m)$ is the prior. The conditional predictive density $Y | (\theta_{i+1}, \dots, \theta_m)$ is defined as the integral within brackets.

⁴This result requires that no data augmentation be needed to approximate the posterior density.

models that require data augmentation to approximate the posterior (e.g., MS VARs, TVP VARs, DFMs, or FAVARs) or other estimators than Chib’s method, such as the Bridge Sampling.

The main findings of the experiment are: *(i)* a fully-computational approach that neglects the *analytical tractability condition* leads to an estimation bias that severely distorts the model ranking when the number of observables is larger than five; *(ii)* both our methods deliver very similar results in terms of posterior model rankings, suggesting that the accuracy of our two methods is of the same order of magnitude in the experiment; *(iii)* exploiting the *analytical tractability condition* prevents our estimators from being affected by the ”curse of dimensionality” (i.e., computing time growing at faster pace as the number of lags and observables in the model increases). In relation to this last finding, we argue that Method 2 is suitable for performing model selection and model averaging across a large number of models.

The paper is organized as follows. Section 2 introduces the conditions that a model has to satisfy in order to apply our two estimators. In this section, we describe the two methods proposed in this paper for computing the MDD. Section 3 discusses the application of our methods to several econometric models. Section 4 performs the Monte Carlo application. Section 5 concludes.

2 Methods for Computing the Marginal Data Density

The marginal data density (MDD), also known as the marginal likelihood of the data, is defined as the integral taken over the likelihood with respect to the prior distribution of the parameters. Let Θ be the parameter set of an econometric model and Y be the sample data. Then, the marginal data density is defined as

$$p(Y) = \int p(Y|\Theta)p(\Theta)d\Theta \tag{1}$$

where $p(Y|\Theta)$ and $p(\Theta)$ denote the likelihood and the prior density, respectively.

In this section, we describe the modeling framework for which we have developed new estimators for the MDD. In particular, we focus on models in which the joint posterior density for parameters can be approximated through the Gibbs sampler. We describe the two methods proposed in this paper in section 2.2.

2.1 The modeling framework

Let us denote a set of observable variables as Y . Let us consider a model whose set of parameters and latent variables is denoted by $\Theta^D = \{D, \Theta\}$ where D stands for the latent variables and Θ for the parameters of the model. We denote the prior for model's parameters as $p(\Theta)$ and it is assumed to have a known analytical representation. Furthermore, the likelihood function, $p(Y|\Theta)$, is assumed to be easy to evaluate. We define blocks in the parameter vector as $\theta_1, \theta_2, \dots, \theta_m$, such that $\Theta \equiv \{\theta_1, \dots, \theta_m\}$. We focus on models whose parameter set, Θ , can be partitioned into at least three parameter blocks (i.e., $m > 2$)⁵ and that satisfy the following two conditions:

- (i) It is possible to draw from the conditional posterior distributions $\theta_i | (\Theta_{-i}, D, Y)$, where $\Theta_{-i} \equiv \{\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_m\}$, for any $i \in \{1, \dots, m\}$ and from the posterior predictive density, $D | (\Theta, Y)$.
- (ii) The conditional posterior distributions $\Theta_{\leq \tau} | (\Theta_{> \tau}, D, Y)$, where $\Theta_{\leq \tau} \equiv \{\theta_1, \dots, \theta_\tau\}$ and $\Theta_{> \tau} \equiv \{\theta_{\tau+1}, \dots, \theta_m\}$, are analytically tractable, for some $\tau \in \{2, \dots, m-1\}$.

Condition (i) implies that we can approximate the joint posterior $\Theta|Y$ and the predictive density $D|Y$ through the Gibbs sampler. We label this condition as the *sampling condition*. Condition (ii) applies when there exists an integer $\tau > 1$ such that the parameter block $\Theta_{< i}$ can be integrated out analytically from the conditional posterior densities $\Theta_{\leq i} | (\Theta_{-i}, D, Y)$ for any $i \in \{2, \dots, \tau\}$. This condition is most likely to be satisfied through a wise partitioning of the parameter set and the specification of a conjugate prior. We refer to this condition as the *analytical tractability condition*. We show that these two conditions are satisfied by a set of models that are widely used in time series and financial econometrics, such as VARs models, Reduced Rank Regression Models, Markov-Switching (MS) VAR models, Time-Varying Parameters (TVP) VARs, Dynamic Factor Models (DFMs), and Factor-augmented VAR models (FAVARs).

2.2 Our Methods

In this section, we present two new methods for computing the marginal data density from the Gibbs output. Method 1 and Method 2 apply to models satisfying both the *sampling*

⁵In most of the models where $m = 2$, the integrating constants of the MDD are available analytically under conjugate priors. Therefore, computing the MDD for these models does not raise any computational issue.

condition and the *analytical tractability condition*.

2.2.1 Method 1

Method 1 is based on interpreting the MDD as the normalizing constant of the joint posterior distribution⁶

$$p(Y) = \frac{p(Y|\Theta)p(\Theta)}{\prod_{j=1}^m p(\theta_j|\Theta_{>j}, Y)} \quad (2)$$

where the numerator is the kernel of the joint posterior and the denominator is the joint posterior distribution. In particular, we focus on settings in which the joint posterior, $p(Y|\Theta)p(\Theta)$, is easy to evaluate.

Factorizing (2) yields

$$\hat{p}_{\text{M1}}(Y) = \hat{p}(Y|\tilde{\Theta}_{>\tau}) \cdot \frac{p(\tilde{\Theta}_{>\tau})}{\hat{p}(\tilde{\Theta}_{>\tau}|Y)} \quad (3)$$

where $\tilde{\Theta}_{>\tau}$ is the parameter set $\Theta_{>\tau} = \{\theta_{\tau+1}, \dots, \theta_m\}$ evaluated at the joint posterior mode $(\tilde{\theta}_j, j \in \{1, \dots, m\})$ and $p(Y|\tilde{\Theta}_{>\tau})$ is the conditional predictive density which is defined as:⁷

$$\hat{p}(Y|\tilde{\Theta}_{>\tau}) = \frac{p(Y|\tilde{\Theta})p(\tilde{\Theta}_{\leq\tau}|\tilde{\Theta}_{>\tau})}{\hat{p}(\tilde{\Theta}_{\leq\tau}|\tilde{\Theta}_{>\tau}, Y)} \quad (4)$$

Method 1 exploits the *analytical tractability condition* and computes:

$$\hat{p}(\tilde{\Theta}_{\leq\tau}|\tilde{\Theta}_{>\tau}, Y) = \frac{1}{n_r} \sum_{s=1}^{n_r} p(\tilde{\Theta}_{\leq\tau}|\tilde{\Theta}_{>\tau}, D^{(s)}, Y) \quad (5)$$

where $\{D^{(s)}\}_{s=1}^{n_r}$ is the output of the reduced Gibbs step (see Algorithm 2 of Appendix A with $i = \tau$) and the conditional posterior $\Theta_{\leq\tau}(\Theta_{>\tau}, D, Y)$ is known because of the *analytical*

⁶We adopt the convention that $\Theta_{>m} = \emptyset$.

⁷To see that this is a component of the MDD, $p(Y)$, decompose the MDD as follows:

$$\begin{aligned} p(Y) &= \int p(Y|\theta_1, \dots, \theta_m) p(\theta_1, \dots, \theta_m) d\theta_1 \dots d\theta_m \\ &= \int \left(\int p(Y|\theta_1, \dots, \theta_m) p(\theta_1, \dots, \theta_\tau | \theta_{\tau+1}, \dots, \theta_m) d\theta_1 \dots d\theta_\tau \right) p(\theta_{\tau+1}, \dots, \theta_m) d\theta_{\tau+1} \dots d\theta_m \end{aligned}$$

where $p(Y|\theta_1, \dots, \theta_m)$ is the likelihood function and $p(\theta_1, \dots, \theta_\tau | \theta_{\tau+1}, \dots, \theta_m) p(\theta_{\tau+1}, \dots, \theta_m)$ is the prior. The conditional predictive density $Y | (\theta_{\tau+1}, \dots, \theta_m)$ is defined in the integral within brackets.

tractability condition. The conditional posteriors $\tilde{\Theta}_{>\tau}|Y$ in (3) are approximated by running $m - \tau - 1$ reduced Gibbs steps (see Algorithm 2 of Appendix A for $i \in \{\tau + 1, m - 1\}$) and using the Rao-Blackwellization technique proposed by Gelfand, Smith, and Lee (1992) to approximate the marginal posterior density $p(\Theta_m)$.⁸

In settings where the *sampling condition* is satisfied, the leading estimator in the literature is the one proposed by Chib (1995). His estimator is also based on interpreting the MDD as the integrating constant of the posterior kernel and relies on Monte Carlo integration. In particular, the only difference with Method 1 is the computation of the conditional density $\hat{p}(\tilde{\Theta}_{\leq\tau}|\tilde{\Theta}_{>\tau}, Y)$ which is given by

$$\hat{p}(\tilde{\Theta}_{\leq\tau}|\tilde{\Theta}_{>\tau}, Y) = \hat{p}(\tilde{\Theta}_1|\tilde{\Theta}_2, \dots, \tilde{\Theta}_m, Y) \hat{p}(\tilde{\Theta}_2|\tilde{\Theta}_3, \dots, \tilde{\Theta}_m, Y) \dots \hat{p}(\tilde{\Theta}_\tau|\tilde{\Theta}_{>\tau}, Y)$$

where only $\hat{p}(\tilde{\Theta}_1|\tilde{\Theta}_2, \dots, \tilde{\Theta}_m, Y)$ is known. The remaining conditional densities are computationally approximated running τ reduced Gibbs steps (i.e., Algorithm 2).

Overall, applying Method 1 requires running $m - \tau$ reduced Gibbs steps as opposed to the $m - 1$ steps performed by Chib’s method. Thus gains from applying Method 1 relative to Chib’s method are expected to become more and more substantial as the number of blocks τ that can be integrated out increases.

If no data augmentation is required by the posterior simulator (i.e., $D = \emptyset$), the *sampling condition* and *analytical tractability condition* imply that a closed-form analytical solution for the conditional predictive density, $p(Y|\tilde{\Theta}_{>\tau})$, in (4) is available. Therefore, Method 1 does not require performing the Monte Carlo integrations in (5) and we only need to run $(m - \tau - 1)$ reduced-Gibbs steps in addition to the Gibbs sampler posterior simulator. In this case, while Chib’s method *computationally approximates* the conditional predictive density $Y|\tilde{\Theta}_{>\tau}$ in (4) via $(m - 2)$ reduced Gibbs steps, Method 1 *exactly calculates* this density using its analytical expression. Therefore, in this setting, Method 1 is, by construction, more accurate and computationally efficient than Chib’s method.

⁸This technique relies on draws from the Gibbs sampler posterior simulator (see Algorithm 1 in Appendix A).

2.2.2 Method 2

Method 2 is based on the principle of Reciprocal Importance Sampling (RIS), proposed by Gelfand and Dey (1994), and stems from observing that

$$\frac{1}{p(Y)} = \mathbb{E}_{p(D, \Theta_{>\tau}|Y)} \left[f(\Theta_{>\tau}) \frac{p(\Theta_{\leq\tau}|\Theta_{>\tau}, D, Y)}{p(Y|\Theta_{\leq\tau}, \Theta_{>\tau}) p(\Theta_{\leq\tau}|\Theta_{>\tau}) p(\Theta_{>\tau})} \right]$$

where $\mathbb{E}_{p(D, \Theta_{>\tau}|Y)}(\cdot)$ denotes the expectations taken with respect to the posterior density $D, \Theta_{>\tau}|Y$.

Method 2 computes the marginal data density, $p(Y)$, as follows:

$$\hat{p}_{M2}(Y) = \left[\frac{1}{n_r} \sum_{s=1}^{n_r} \frac{p(\tilde{\Theta}_{\leq\tau}|\Theta_{>\tau}^{(s)}, D^{(s)}, Y)}{p(Y|\tilde{\Theta}_{\leq\tau}, \Theta_{>\tau}^{(s)}) p(\tilde{\Theta}_{\leq\tau}|\Theta_{>\tau}^{(s)}) p(\Theta_{>\tau}^{(s)})} f(\Theta_{>\tau}^{(s)}) \right]^{-1} \quad (6)$$

where $\{\Theta_{>\tau}^{(s)}, D^{(s)}\}$ are the draws from the Gibbs sampler simulator (Algorithm 1 in Appendix A), $\tilde{\Theta}_{\leq\tau}$ denotes the posterior mode and $f(\cdot)$ is a weighting function, such as $\int f(\Theta_{>\tau}) d\Theta_{>\tau} = 1$. The numerator is the conditional posterior, which is known because of the *analytical tractability condition*. In the denominator, we have that the likelihood function $p(Y|\tilde{\Theta}_{\leq\tau}, \Theta_{>\tau}^{(s)})$ and the joint prior $p(\tilde{\Theta}_{\leq\tau}|\Theta_{>\tau}^{(s)}) p(\Theta_{>\tau}^{(s)})$ are evaluated at the posterior mode for $\Theta_{\leq\tau}$ and at the s -th Gibbs sampler draw for $\Theta_{>\tau}$. It should be noticed that, unlike Method 1 and Chib's Method, Method 2 is a hybrid estimator that evaluates the densities in (6) locally for the parameter blocks $\Theta_{\leq\tau}$ and globally for the parameter blocks in $\Theta_{>\tau}$.

If no data augmentation (i.e., $D = \emptyset$) is required, then the *analytical tractability condition* implies that the analytical expression of the conditional predictive density $p(Y|\Theta_{>\tau})$ defined as

$$p(Y|\Theta_{>\tau}) = \frac{p(Y|\Theta) p(\Theta_{\leq\tau}|\Theta_{>\tau})}{p(\Theta_{\leq\tau}|\Theta_{>\tau}, Y)} \quad (7)$$

is available. Hence, equation (6) simplifies to:

$$\hat{p}_{M2}(Y) = \left[\frac{1}{n} \sum_{s=1}^n \frac{f(\Theta_{>\tau}^{(s)})}{p(Y|\Theta_{>\tau}^{(s)}) p(\Theta_{>\tau}^{(s)})} \right]^{-1} \quad (8)$$

where $p(Y|\Theta_{>\tau}^{(s)})$ is the conditional predictive density defined in (7) and the draws $\Theta_{>\tau}^{(s)}$ are the draws from the Gibbs sampler simulator. Thus, when data augmentation is not

necessary, Method 2 becomes a global estimator.

In this paper, we follow Geweke (1999) and define the weighting function $f(\Theta_{>\tau}^{(s)})$ as

$$f(\Theta_{>\tau}^{(s)}) = \frac{1}{\tau} (2\pi)^{-d/2} |V|^{-1/2} \exp\left\{-\frac{1}{2}(\Theta_{>\tau}^{(s)} - \tilde{\Theta}_{>\tau})' V^{-1} (\Theta_{>\tau}^{(s)} - \tilde{\Theta}_{>\tau})\right\} \\ \times \mathfrak{I}\{(\Theta_{>\tau}^{(s)} - \tilde{\Theta}_{>\tau})' V^{-1} (\Theta_{>\tau}^{(s)} - \tilde{\Theta}_{>\tau}) \leq F_{\chi_d^2(\nu)}\} \quad (9)$$

where d is the dimension of the parameter vector $vec(\Theta_{>\tau})$, $\mathfrak{I}\{(\Theta_{>\tau}^{(s)} - \tilde{\Theta}_{>\tau})' V^{-1} (\Theta_{>\tau}^{(s)} - \tilde{\Theta}_{>\tau}) \leq F_{\chi_d^2(\nu)}\}$ is an indicator function, and $F_{\chi_d^2(\nu)}$ is the cumulative distribution function of a chi-square distribution with ν degrees of freedom. The hyperparameter ν has to be chosen so as to minimize the numerical standard error of the estimator. It is important to emphasize that fine-tuning this parameter does not require performing again any Gibbs sampler or any additional evaluations of densities or functions.

To sum up, Method 1 overlaps Chib's method when performing reduced Gibbs steps for $i \in \{\tau + 1, m - 1\}$. Note that these simulations are the most computationally cumbersome among all the reduced Gibbs steps performed by Chib's method as they are the ones which integrate out the largest number of parameter blocks. When the total number of parameter blocks, m , is much larger than the number of blocks τ that can be integrated out, then Method 1 may be still computationally cumbersome. In these cases and when a large number of repeated computations of MDDs is required (e.g., Bayesian averaging over a large number of models), Method 2 provides the fastest approach. It is important to emphasize that Method 2 only requires performing the Gibbs sampler posterior simulator, regardless of the number of partitions that can be integrated out, τ , and whether data augmentation is required.

3 A Guide to Apply Method 1 and Method 2

In this section, we define the set of models, to which our methods are applicable. This set includes models that are very popular in time series econometrics, such as VAR models, Reduced Rank Regression (RRR) models, which include, for instance, Vector Equilibrium Correction models (VECM), Markov-switching VAR models, Time-varying parameters VAR models, Dynamic Factor Models (DFMs), and Factor Augmented VAR models (FAVARs).

3.1 Vector Autoregressive Models

Following Villani (2009) and Del Negro and Schorfheide (2010), we parameterize the VAR model in mean-adjusted form

$$y_t = \sum_{j=0}^l \gamma_j t^j + \tilde{y}_t \quad (10)$$

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \dots + \phi_p \tilde{y}_{t-p} + \varepsilon_t, \quad \varepsilon_t \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma) \quad (11)$$

where γ_j , $j \in \{0, \dots, l\}$, and \tilde{y}_t are $n \times 1$ vectors. The sum $\sum_{j=0}^l \gamma_j t^j$ captures the deterministic trend and \tilde{y}_t the stochastic fluctuations around it. This specification is flexible enough to capture any representation for a VAR model. The mean-adjusted representation not only encompasses models with any deterministic trend (linear, quadratic, cubic, etc.), but also models with stochastic-trends⁹. As pointed out by Villani (2009), parameterizing the VAR model as in (10)-(11) makes it straightforward to separate beliefs about the deterministic trend component from beliefs about the persistence of fluctuations around this trend.

We can recast the model (10)-(11) in matrix notation as

$$Y = D\Gamma + \tilde{Y} \quad (12)$$

$$\tilde{Y} = \tilde{X}\Phi + \varepsilon \quad (13)$$

where we denote the sample length as T and we define the $T \times n$ matrix $Y = (y_1, \dots, y_T)'$, the $T \times (l+1)$ matrix $D = \left[\mathbf{1}'_T, (1, \dots, T)', \dots, (1, \dots, T^l)' \right]$ with $\mathbf{1}_T$ being a $1 \times T$ vector of ones, the $(l+1) \times n$ matrix $\Gamma = (\gamma_0, \dots, \gamma_l)'$, the $T \times n$ matrix \tilde{Y} is defined as $\tilde{Y} = (\tilde{y}_1, \dots, \tilde{y}_T)'$, the $T \times np$ matrix $\tilde{X} = (\tilde{x}_1, \dots, \tilde{x}_T)'$, where we define the $np \times 1$ vectors $\tilde{x}_t = (\tilde{y}'_{t-1}, \dots, \tilde{y}'_{t-p})'$, the $np \times n$ parameter matrix $\Phi = [\phi_1, \dots, \phi_p]'$, and the $T \times n$ matrix of residuals is denoted as $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)'$.

Let us block partition the parameter space into the following three blocks: $\theta_1 = \Phi$, $\theta_2 = \Sigma$, and $\theta_3 = \Gamma$. Note that, conditional on the parameter block Γ , the equations (12)-(13) can be interpreted as a Multivariate Linear Gaussian Regression Model. Therefore, the posterior distribution $(\Phi, \Sigma) | \Gamma, Y$ is conjugate and analytically tractable belonging to the Multivariate-Normal-Inverted-Wishart ($\mathcal{MN}\mathcal{IW}$) family. This suffices to guarantee the

⁹We can obtain a stochastic-trend model by simply setting $\gamma_j = 0$ for $j = 1, \dots, l$.

satisfaction of the *analytical tractability condition* for $\tau = 2$. Moreover, if the prior for Γ is independent and Gaussian, the conditional posterior $\Gamma | (\Phi, \Sigma, Y)$ can be shown to be also Gaussian (see the Appendix C). Therefore, the *sampling condition* is satisfied.

For these models, a class of conjugate priors $p(\Phi, \Sigma | \Gamma)$ of the \mathcal{MNTW} family can be obtained through dummy-observation priors. This class of priors is very broad and include widely used prior densities: (a) the Minnesota prior proposed by Litterman (1980) which is based on the assumption that each of the components of y_t is best represented by a random walk process; (b) the sum-of-coefficients prior proposed by Doan, Litterman, and Sims (1984); (c) the single-unit-root prior proposed by Sims and Zha (1998); and (d) priors from macroeconomic models introduced by Del Negro and Schorfheide (2004).

In this context, the estimator for the marginal data density proposed by Chib (1995) is given by:

$$\hat{p}_{\text{CHIB}}(Y) = \frac{p(Y | \tilde{\Sigma}, \tilde{\Phi}, \tilde{\Gamma}) p(\tilde{\Sigma}, \tilde{\Phi}, \tilde{\Gamma})}{p(\tilde{\Phi} | \tilde{\Sigma}, \tilde{\Gamma}, Y) \hat{p}(\tilde{\Sigma} | \tilde{\Gamma}, Y) \hat{p}(\tilde{\Gamma} | Y)} \quad (14)$$

where the triplet $(\tilde{\Sigma}, \tilde{\Phi}, \tilde{\Gamma})$ stands for the mode of the joint posterior density $p(\Phi, \Sigma, \Gamma | Y)$. The numerator of (14) is the posterior kernel conveniently factorized and the denominator is the joint posterior.

Given that (i) the *sampling condition* is satisfied, (ii) data augmentation is not required, and (iii) the likelihood function of the model (12)-(13) is available in closed-form expression,¹⁰ we can evaluate all the terms in (14) with the exception of $p(\tilde{\Sigma} | \tilde{\Gamma}, Y)$ and $p(\tilde{\Gamma} | Y)$. Chib (1995) suggests to evaluate the marginal posterior $p(\tilde{\Gamma} | Y)$ implementing a Rao-Blackwell strategy that uses the output from the Gibbs sampler as follows:

$$\hat{p}(\tilde{\Gamma} | Y) \approx \frac{1}{n} \sum_{s=1}^n p(\tilde{\Gamma} | \Sigma^{(s)}, \Phi^{(s)}, Y) \quad (15)$$

To compute the density $p(\tilde{\Sigma} | \tilde{\Gamma}, Y)$, Chib's method performs one reduced Gibbs step (see Algorithm 2 with $i = 2$ in Appendix A) For the sake of clarity, we detail the reduced Gibbs steps for VAR models below.

Algorithm 3: Reduced-Gibbs Sampler for VAR Models:

Given an initial set of parameter values, $\{\Sigma^{(0)}, \Phi^{(0)}\}$ set $s = 0$ and perform the following steps:

¹⁰We consider the conditional likelihood on the first initial observations. We keep this assumption throughout the paper.

1. Draw $\Sigma^{(s+1)}$ from $p\left(\Sigma|\tilde{\Gamma}, Y\right)$.
2. Draw $\Phi^{(s+1)}$ from $p\left(\Phi|\Sigma^{(s+1)}, \tilde{\Gamma}, Y\right)$.
3. Set $s = s + 1$. If $s \leq n_r$, go to step 1. Otherwise STOP.

The output from the reduced Gibbs step can be used to computationally evaluate to $p\left(\tilde{\Sigma}|\tilde{\Gamma}, Y\right)$ as follows:

$$\hat{p}\left(\tilde{\Sigma}|\tilde{\Gamma}, Y\right) \approx \frac{1}{m} \sum_{s=1}^m p\left(\tilde{\Sigma}|\Phi^{(s)}, \tilde{\Gamma}, Y\right) \quad (16)$$

Method 1 computes:

$$\hat{p}_{\text{M1}}(Y) = p\left(Y|\tilde{\Gamma}\right) \cdot \frac{p\left(\tilde{\Gamma}\right)}{\hat{p}\left(\tilde{\Gamma}|Y\right)} \quad (17)$$

Since the *sampling* and *analytical tractability conditions* (with $\tau = 2$) are satisfied, no data augmentation is required, and the likelihood function of the model (12)-(13) is available in closed-form expression, the conditional predictive density, $p\left(Y|\tilde{\Gamma}\right)$, in (17), has an analytical closed form solution:¹¹

$$p\left(Y|\tilde{\Gamma}\right) = \frac{\pi^{-\frac{(T_0+T_1-np)n}{2}} \left|\bar{X}'\bar{X}\right|^{-\frac{n}{2}} |S|^{-\frac{T_0+T_1-np}{2}} \cdot \Gamma_n\left(\frac{T_0+T_1-np}{2}\right)}{\pi^{-\frac{(T_0-np)n}{2}} \left|X^{*'}X^*\right|^{-\frac{n}{2}} |S^*|^{-\frac{T_0-np}{2}} \cdot \Gamma_n\left(\frac{T_0-np}{2}\right)} \quad (18)$$

where Y^* and X^* are the dummy observations for the VAR in deviations. T_0 is the number of dummy observations, $T_1 = T + T_0$, $\bar{Y} = \left[Y^{*'}, \tilde{Y}'\right]'$, $\bar{X} = \left[X^{*'}, \tilde{X}'\right]'$, $\Gamma_n(\cdot)$ is the multivariate gamma function, $\mathbf{S} = \left(\bar{Y} - \bar{X}\hat{\Phi}\right)' \left(\bar{Y} - \bar{X}\hat{\Phi}\right)$ with $\hat{\Phi} = \left(\bar{X}'\bar{X}\right)^{-1} \bar{X}'\bar{Y}$ and $\mathbf{S} = \left(Y^* - X^*\hat{\Phi}^*\right)' \left(Y^* - X^*\hat{\Phi}^*\right)$ with $\hat{\Phi}^* = \left(\bar{X}'\bar{X}\right)^{-1} \bar{X}'\bar{Y}$. Furthermore, Method 1 estimates the marginalized posterior $\hat{p}\left(\tilde{\Gamma}|Y\right)$ as Chib's method in (15). A detailed derivation of the conditional predictive density $Y|\tilde{\Gamma}$ is provided in the Appendix D.

A naïve application of Chib's method disregards the formula in (18) and computationally approximates the conditional predictive density by calculating

$$\hat{p}_{\text{CHIB}}(Y|\tilde{\Gamma}) = \frac{p(Y|\tilde{\Sigma}, \tilde{\Phi}, \tilde{\Gamma})p(\tilde{\Sigma}, \tilde{\Phi}|\tilde{\Gamma})}{p(\tilde{\Phi}|\tilde{\Sigma}, \tilde{\Gamma}, Y)p(\tilde{\Sigma}|\tilde{\Gamma}, Y)} \quad (19)$$

¹¹For a derivation of this formula see Zellner (1971).

where $p(\tilde{\Sigma}|\tilde{\Gamma}, Y)$ is approximated as in (16) through the output of the reduced-Gibbs step in Algorithm 3. In contrast, Method 1 takes a fully analytical approach and exactly calculates the conditional predictive density $p(Y|\tilde{\Gamma})$ via its formula in equation (18). Thus, by construction, Method 1 is more accurate and less computationally burdensome than Chib’s method in the context of mean-adjusted VARs.

To apply Method 1, we only need the draws from the Gibbs sampler posterior simulator in order to evaluate the density $p(\tilde{\Gamma}|Y)$ in (17).

Method 2 computes:

$$\hat{p}_{M2}(Y) = \left[\frac{1}{n} \sum_{s=1}^n \frac{f(\Gamma^{(s)})}{p(Y|\Gamma^{(s)})p(\Gamma^{(s)})} \right]^{-1} \quad (20)$$

where the draws $\Gamma^{(s)}$ are the draws from the Gibbs sampler simulator¹². We analytically evaluate the posterior kernel $p(Y|\Gamma^{(s)})p(\Gamma^{(s)})$ and the weighting function $f(\Gamma^{(s)})$ for each draw of Γ .

3.2 Reduced Rank Regression Models

Bayesian analysis of reduced rank regression (RRR) models is detailed in Geweke (1996). The RRR model reads:

$$Y = X\Gamma + Z\Phi + u_t \quad (21)$$

with $u_t \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma)$. X and Z are $n \times p$ and $n \times k$ matrices of explanatory variables, respectively. Γ and Φ are matrices of unknown coefficients, whose dimensions are $p \times L$ and $k \times L$, respectively. While the matrix of coefficients, Φ is full-rank, the matrix Γ , is assumed to have rank q , where $q < \max\{L, p\}$. Since Γ is a low-rank matrix, we can reparameterized it as $\Gamma = \Psi\Omega$. It is important to emphasize that the matrices Ψ and Ω cannot be identified under an improper, flat prior distribution for Ψ and Ω . We need to resort to some normalization to identify these matrices. Normalization schemes are applied to either matrix Ψ or Ω . As a result, there are two classes of normalization schemes. First, schemes that restrict Ω to be Ω^* (normalization 1). Second, schemes that restrict Ψ to be Ψ^* (normalization 2).¹³ In the remaining of this section, we focus on RRR models where the

¹²In order to implement this approach, we need the draws $\{\Gamma\}_{s=1}^n$ from the marginalized posterior $p(\Gamma|Y)$. It can be shown that these draws are simply the set of draws $\{\Gamma\}_{s=1}^n$ that come from the output of the Gibbs sampler.

¹³Popular normalizations are discussed in Del Negro and Schorfheide (2010). See also Strachan and Inder (2004) for a critical viewpoint of ordinal normalization schemes.

low-rank matrix Γ is identified through normalization 2. It is straightforward to extend the results to normalization 1.

Geweke (1996) proposes as a reference prior the product of an independent inverted Wishart distribution for Σ and independent Gaussian shrinkage priors for each of the elements of Ψ^* and Ω . Let us partition the parameter space of the RRR model in (21) as follows $\theta_1 = \Phi$, $\theta_2 = \Sigma$, $\theta_3 = \Psi^*$, and $\theta_4 = \Omega$. Geweke (1996) shows that the conditional predictive densities $\Phi | (\Sigma, \Psi^*, \Omega, Y)$, $\Sigma | (\Psi^*, \Omega, Y)$, $\Psi^* | (\Phi, \Sigma, \Omega, Y)$, and $\Omega | (\Phi, \Sigma, \Psi^*, Y)$ belong to the $\mathcal{MN}\mathcal{TW}$ family. Therefore, the *sampling condition* is satisfied. Note that, conditional on Γ , the RRR model in (21) reduces to a multivariate linear Gaussian regression model. Given a $\mathcal{MN}\mathcal{TW}$ prior on $(\Phi, \Sigma) | \Gamma$, we conclude that the posterior $(\Phi, \Sigma) | (\Gamma, Y)$ is not only $\mathcal{MN}\mathcal{TW}$ but also analytically tractable. Hence, the *analytical tractability condition* is satisfied for $\tau = 2^{14}$.

Chib's method computes:

$$\hat{p}_{\text{CHIB}}(Y) = \frac{p(Y|\tilde{\Sigma}, \tilde{\Phi}, \tilde{\Psi}^*, \tilde{\Omega})p(\tilde{\Sigma}, \tilde{\Phi}, \tilde{\Gamma}, \tilde{\Psi}^*, \tilde{\Omega})}{p(\tilde{\Phi}|\tilde{\Sigma}, \tilde{\Psi}^*, \tilde{\Omega}, Y)\hat{p}(\tilde{\Sigma}|\tilde{\Psi}^*, \tilde{\Omega}, Y)\hat{p}(\tilde{\Psi}^*|\tilde{\Omega}, Y)\hat{p}(\tilde{\Omega}|Y)} \quad (22)$$

where $(\tilde{\Sigma}, \tilde{\Phi}, \tilde{\Psi}^*, \tilde{\Omega})$ stands for the mode of the joint posterior density $(\Phi, \Sigma, \Psi^*, \Omega) | Y$. Chib's method needs to perform two reduced Gibbs steps in addition to the Gibbs sampler to approximate $\tilde{\Sigma} | (\tilde{\Psi}^*, \tilde{\Omega}, Y)$, $\tilde{\Psi}^* | (\tilde{\Omega}, Y)$ and $\tilde{\Omega} | Y$.

Method 1 computes:

$$\hat{p}_{\text{M1}}(Y) = p(Y|\tilde{\Psi}^*, \tilde{\Omega}) \cdot \frac{p(\tilde{\Psi}^*, \tilde{\Omega})}{\hat{p}(\tilde{\Psi}^*|\tilde{\Omega}, Y)\hat{p}(\tilde{\Omega}|Y)} \quad (23)$$

where the conditional predictive density, $p(Y|\Psi^*, \Omega)$, can be shown to have the following analytical closed-form expression:

$$p(Y|\tilde{\Psi}^*, \tilde{\Omega}) = \pi^{-\frac{L(n-k)}{2}} |Z'Z|^{-\frac{L}{2}} |S|^{-\frac{n-k}{2}} \pi^{\frac{L(L-1)}{4}} \prod_{i=1}^L \Gamma[(n-k+1-i)/2] \quad (24)$$

where $\Gamma_L(\frac{n-k}{2})$ is the multivariate gamma function and

$$S \equiv (Y - X\tilde{\Psi}^*\tilde{\Omega} - Z\tilde{\Phi})' (Y - X\tilde{\Psi}^*\tilde{\Omega} - Z\tilde{\Phi})$$

¹⁴This result still holds if we consider an improper prior, such that $p(\Phi, \Sigma, \Gamma) \propto |\Sigma|^{-(L+1)/2}$.

$\hat{\Phi} \equiv (Z'Z)^{-1} Z' (Y - X\tilde{\Psi}^*\tilde{\Omega})$.¹⁵ The densities $\tilde{\Psi}^* | (\tilde{\Omega}, Y)$ and $\tilde{\Omega} | Y$ are estimated exactly as in Chib's method. Overall, Method 1 requires performing only one reduced-Gibbs step in addition to the Gibbs sampler.

Method 2 computes:

$$\hat{p}_{M2}(Y) = \left[\frac{1}{n} \sum_{s=1}^n \frac{f(\Psi^{*(s)}, \Omega^{(s)})}{p(Y|\Psi^{*(s)}, \Omega^{(s)})p(\Psi^{*(s)}, \Omega^{(s)})} \right]^{-1} \quad (25)$$

where the draws $(\Psi^{*(s)}, \Omega^{(s)})$ are the draws from the Gibbs sampler simulator. We analytically evaluate the density $p(Y|\Psi^{*(s)}, \Omega^{(s)})$, the prior $p(\Psi^{*(s)}, \Omega^{(s)})$, and the weighting function $f(\Psi^{*(s)}, \Omega^{(s)})$ for each draw $(\Psi^{*(s)}, \Omega^{(s)})$. Note that Method 2 does not require any reduced-Gibbs step to be implemented.

A particular class of RRR models is the Vector Error Correction Model (VECM). These models have been applied to study a wide range of issues in time series and financial econometrics.¹⁶ This is only a particular reparameterization of reduced-form VAR models, usually undertaken when the observables, Y , have a unit root but there are linear combinations of observables (i.e., Ω^*y_t) that are stationary.

3.3 Markov-Switching (MS) VARs

Markov-Switching Vector Autoregressive Models (MS-VAR), popularized by Hamilton (1989), are used to capture sudden changes in time-series dynamics. In particular, MS-VAR models are specified such that the coefficients of the reduced form VAR are subject to regime switching.

$$y'_t = x'_t \Phi(K_t) + u'_t \quad (26)$$

where y_t is a $n \times 1$ vector of observable variables, $x'_t = [y'_{t-1}, \dots, y'_{t-p}, 1]$, $\Phi(K_t) = [\Phi_1(K_t), \dots, \Phi_p(K_t), \Phi_c(K_t)]'$, and $u_t \sim N(0, \Sigma(K_t))$. K_t is a discrete M -state Markov process with time-invariant transition probabilities:

$$\pi_{lm} = P[K_t = l | K_{t-1} = m], \quad l, m \in \{1, \dots, M\} \quad (27)$$

¹⁵Derivation of expression (87) is straightforward given that, conditional on Γ , the RRR model in (21) reduces to a multivariate linear Gaussian regression model. So the exact analytical form for the conditional predictive density $Y|\tilde{\Psi}^*, \tilde{\Omega}$ can be obtained along the lines of the discussion presented in Appendix D. See also Zellner (1971).

¹⁶A useful survey is provided by Koop, Strachan, van Dijk, and Villani (2006).

For simplicity, let us assume that $M = 2$. Let T be the sample length, $K = (K_1, \dots, K_T)$ be the history of regimes, $[\Phi(j), \Sigma(j)]_{j \in \{1,2\}} = \{\Phi(1), \Sigma(1), \Phi(2), \Sigma(2)\}$, and $(\pi_{jj})_{j \in \{1,2\}} = \{\pi_{11}, \pi_{22}\}$. It is convenient to partition the parameter space of the MS-VAR model in (26)-(27) as follows $\theta_1 = (\pi_{jj})_{j \in \{1,2\}}$, $\theta_2 = \Phi(1)$, $\theta_3 = \Sigma(1)$, $\theta_4 = \Phi(2)$, $\theta_5 = \Sigma(2)$.

Conditional on the history of regimes, K , (i) the model (26)-(27) reduces to a VAR model with dummy variables that account for known structural breaks and (ii) the transition probabilities, $(\pi_{jj})_{j \in \{1,2\}}$, are independent of the data and of the remaining parameters of the model, $[\Phi(j), \Sigma(j)]_{j \in \{1,2\}}$. As a result, if the prior distributions for $\Phi(l)$ and $\Sigma(l)$, $l \in \{1,2\}$, are of the $\mathcal{MN}\mathcal{TW}$ form and π_{11} and π_{22} are independent beta distributions, then conditional posterior distributions of $(\Phi(l), \Sigma(l)) | (K, Y)$, $l \in \{1,2\}$ and $(\pi_u, l = 1, 2) | (Y, K, (\Phi(j), \Sigma(j))_{j \in \{1,2\}})$ belong to the same family of their corresponding priors. Therefore, the *analytical tractability condition* is satisfied for $\tau = 5$.¹⁷ Since the draws from the conditional posterior distribution for the regimes $K | (Y, (\Phi(j), \Sigma(j))_{j \in \{1,2\}}, (\pi_u, l = 1, 2))$ can be obtained using a variant of the Carter and Kohn (1994)¹⁸, the *sampling condition* is also satisfied.

The application of the Chib's method is straightforward, so we do not discuss it here. Given that $\tau = m$, equation (3), which characterizes Method 1, reduces to

$$\hat{p}_{M1}(Y) = \frac{p(Y|\tilde{\Theta})p(\tilde{\Theta})}{\hat{p}(\tilde{\Theta}|Y)} \quad (28)$$

where $\tilde{\Theta} \equiv \left[(\tilde{\pi}_{jj})_{j \in \{1,2\}}, (\tilde{\Phi}(j), \tilde{\Sigma}(j))_{j \in \{1,2\}} \right]$ is the vector of posterior modes. The likelihood $p(Y|\tilde{\Theta})$ does not have a closed-form solution but it can be easily evaluated using the expectation-maximization approach discussed in Kim and Nelson (1999) (chapter 10). Method 1 approximates the joint posterior density $p(\tilde{\Theta}|Y)$ from the output of the Gibbs sampler as follows:

$$\hat{p}(\tilde{\Theta}|Y) = \frac{1}{n_r} \sum_{s=1}^{n_r} \prod_{j=1}^2 p(\tilde{\pi}_{jj}|K^{(s)}, Y) \cdot p(\tilde{\Phi}(j), \tilde{\Sigma}(j)|K^{(s)}, Y) \quad (29)$$

where $p(\pi_{11}|K^{(s)}, Y) \sim \text{Beta}(\cdot)$, $p(\pi_{22}|K^{(s)}, Y) \sim \text{Beta}(\cdot)$ and $p(\Phi(j), \Sigma(j)|K^{(s)}, Y) \sim$

¹⁷These restrictions over priors are only sufficient for satisfying the *analytical tractability condition*. Such condition can be shown to be also satisfied under an improper flat prior, such as $\prod_{j=1}^2 p(\Phi(j), \Sigma(j)) = \prod_{j=1}^2 |\Sigma(j)|^{-(n+1)/2}$.

¹⁸See Del Negro and Schorfheide (2010) and Pitt and Kohn (2010)

$MNIW(\cdot)$. The exact formula for these three density is shown in the Appendix E.

Method 2 computes:

$$\hat{p}_{M2}(Y) = \left[\frac{1}{n_r} \sum_{s=1}^{n_r} \frac{p(\tilde{\Theta}|K^{(s)}, Y)}{p(Y|\tilde{\Theta}) p(\tilde{\Theta})} \right]^{-1}$$

where $K^{(s)}$ are the n_r posterior draws obtained from the multimove Gibbs sampler proposed by Carter and Kohn (1994). All the densities on the right-hand side have a known analytical characterization except for the likelihood $p(Y|\tilde{\Theta})$.

A naive application of Chib's method would lead to perform four reduced Gibbs steps in addition to the Gibbs sampler. Hence, gains in computing time from Method 1 and Method 2 are expected to be large, since these methods only require using the draws from the Gibbs sampler posterior simulator. It is worthwhile emphasizing that while generating draws from the Gibbs sampler are necessary for Bayesian inference, draws from the reduced Gibbs step have much more limited utility in standard applications.

3.4 Time-Varying Parameters (TVP) VAR Models

VAR models with time-varying coefficients have become popular in macroeconometrics since the papers by Cogley and Sargent (2002, 2005) and Primiceri (2005).

Following the notation in Primiceri (2005), a TVP VAR model is given by

$$y_t = X_t' \phi_t + u_t \tag{30}$$

where the $n \times 1$ vector y_t includes the observable variables at time t , the $(np + 1) \times 1$ vectors $x_t = (1, y_{t-1}', \dots, y_{t-p}')'$, the $T \times (np + 1)$ matrix $X_t = \mathbb{I}_n \otimes x_t$, and the $n \times 1$ vector u_t includes the shocks at time t . The vector of parameters, ϕ_t , is assumed to evolve according to a random walk process

$$\phi_t = \phi_{t-1} + \nu_t, \quad \nu_t \sim \mathcal{N}(0, Q) \tag{31}$$

It is standard to restrict the covariance matrix Q to be diagonal and the parameter innovations, ν_t , to be uncorrelated with the VAR innovations, u_t . Furthermore, we assume that

the u_t innovations are Gaussian with heteroskedastic variance:

$$u_t \sim \mathcal{N}(0, \Sigma_t), \quad \Sigma_t = B_t^{-1} H_t B_t^{-1} \quad (32)$$

In the decomposition of Σ_t , the matrix B_t is a lower-triangular matrix with unitary diagonal elements. The vector collecting the non-zero and off-diagonal elements of the matrix B_t evolves as a random walk

$$\alpha_t = \alpha_{t-1} + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, S) \quad (33)$$

Finally, the time-varying matrix H_t is diagonal with elements $h_{i,t}^2$, $i \in \{1, \dots, n\}$, following the geometric random walk:

$$\ln h_t = \ln h_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, W) \quad (34)$$

where the $n \times 1$ vector $h_t = (h_{1,t}, \dots, h_{n,t})'$. Matrices Q , S , and W are positive-definite matrices.

The latent variables of the TVP VAR model (30)-(34) are $D_{0:t} = (\phi_{0:t}, \alpha_{0:t}, \ln h_{0:t})$ and its parameter set is $\Theta = (\phi_0, \alpha_0, \ln h_0, Q, S, W)$, where the first three elements are the initial values for the latent variables.

We partition the parameter space Θ as follows $\theta_1 = \phi_0$, $\theta_2 = \alpha_0$, $\theta_3 = \ln h_0$, $\theta_4 = \text{vec}(Q)$, $\theta_5 = \text{vec}(S)$, $\theta_6 = \text{vec}(W)$. Following Primiceri (2005), we use conjugate priors. In particular, we consider independent Gaussian priors for the initial conditions and independent inverted Wishart priors for the covariance matrices. It directly follows that the joint posterior can be written as

$$p(\Theta | D_{0:t}, Y) = p(\phi_0 | \phi_{0:T}) p(\alpha_0 | \alpha_{0:T}) p(\ln h_0 | \ln h_{0:T}) p(Q | \phi_{0:T}) p(S | \alpha_{0:T}) p(W | \ln h_{0:T}) \quad (35)$$

where all the densities on the right-hand side are known. This implies that the *analytical tractability condition* is satisfied for $\tau = 6$. Primiceri (2005) shows that the *sampling condition* is also satisfied.

The application of the Chib's method is pretty straightforward requiring five reduced-Gibbs steps. Method 1 is performed according to the formula in (28). The likelihood $p(Y | \tilde{\Theta})$ does not have a closed-form solution but it can be evaluated through the Kalman filter, as shown by Primiceri (2005). Method 1 approximates the joint posterior density $p(\tilde{\Theta} | Y)$ from

the output of the Gibbs sampler as follows:

$$\hat{p}(\tilde{\Theta}|Y) = \frac{1}{n_r} \sum_{s=1}^{n_r} p(\tilde{\Theta}|D_{0:t}, Y) \quad (36)$$

where $p(\tilde{\Theta}|D_{0:t}, Y)$ is defined in (35) and the draws $D_{0:t}$ are the draws obtained from the Gibbs sampler posterior simulator (for details, see Primiceri (2005)).

Method 2 computes:

$$\hat{p}_{M2}(Y) = \left[\frac{1}{n_r} \sum_{s=1}^{n_r} \frac{p(\tilde{\Theta}|D_{0:t}, Y)}{p(Y|\tilde{\Theta}) p(\tilde{\Theta})} \right]^{-1}$$

All the densities on the right-hand side have a known analytical characterization, except for the likelihood, which can be evaluated through the Kalman filter.

The inaccuracy and the computational burden associated with naively applying fully computational estimators, such as Chib's estimator, is expected to be large. Our methods are a step forward in trying to reduce the burden since they only require draws from the Gibbs sampler.

3.5 Dynamic Factor Models

Over the last decades, empirical macroeconomists have relied on factor models to analyze the dynamics of time series separating common components from idiosyncratic ones. Factor models in macroeconomics were first studied by Geweke (1977) and Sargent and Sims (1977), but popularized by Stock and Watson (1989). The rising popularity of dynamic factor models (DFM) has been linked to their ability to summarize efficiently large data sets.¹⁹

DFMs decompose the behavior of n observable variables $y_{i,t}$, $i = 1, \dots, n$, into the sum of two unobservable components: for any $t = 1, \dots, T$,

$$y_{i,t} = a_i + \lambda_i f_t + \xi_{i,t}, \quad \xi_{i,t} \overset{iid}{\sim} \mathcal{N}(0, \sigma_i^2) \quad (37)$$

¹⁹Among the multiple applications of DFMs, we can highlight their use in the construction of coincident and leading indicators by Stock and Watson (1989); in forecasting time series by Stock and Watson (1999, 2002a, 2002b), Forni, Hallin, Lippi, and Reichlin (2003), and Boivin and Ng (2005); in real-time monitoring by Giannone, Reichlin, and Small (2008) and Aruoba, Diebold, and Scotti (2009); and in the study of international business cycles by Kose, Otrok, and Whiteman (2003, 2008).

where a_i is a constant; f_t is a $k \times 1$ vector of factors which are common to all observables, λ_i is a $1 \times k$ vector of loadings that links the observables to the factors, and $\xi_{i,t}$ is an innovation specific to each observable variable. The factors evolve according to a vector autoregressive process:

$$f_t = \Phi_{0,1}f_{t-1} + \dots + \Phi_{0,q}f_{t-q} + u_{0,t}, \quad u_{0,t} \overset{iid}{\sim} \mathcal{N}(0, \Sigma_0) \quad (38)$$

where $u_{0,t}$ is a $k \times 1$ vector and the matrices $\Phi_{0,j \in \{1, \dots, p\}}$ and Σ_0 are $k \times k$ matrices. The stochastic vector of innovations, u_t , has dimension of $k \times 1$.

The key assumption is that, at all leads and lags, the $\xi_{i,t}$ innovations are independent across i and independent of the innovations to the factors, $u_{0,t}$. This assumption helps identifying the factor model in (38) by implying that all co-movements in the data arise through the factors. Nonetheless, the factors and the coefficients matrices of the factor model in (38) cannot be identified unless further restrictions are imposed. A popular approach is to impose restrictions upon the variance-covariance matrix of the factor model, Σ_0 , and on the first k loadings, $\lambda_1, \dots, \lambda_k$. See, for instance, Geweke and Zhou (1996) and Del Negro and Schorfheide (2010). We denote the restricted matrix Σ_0 as Σ_0^* and the restricted matrix of factor loadings as $\lambda^* = (\lambda_1^*, \dots, \lambda_k^*, \dots, \lambda_n)'$.

Let us define the $n \times 1$ vectors $y_t = (y_{1,t}, \dots, y_{n,t})'$, $a = (a_1, \dots, a_n)'$, $\lambda^* = (\lambda_1^*, \dots, \lambda_k^*, \dots, \lambda_n)'$, $\xi_t = (\xi_{1,t}, \dots, \xi_{n,t})'$, and, for any $j \in \{1, \dots, p\}$, the $n \times n$ diagonal matrix Φ_j , whose diagonal elements are $(\phi_{1,j}, \dots, \phi_{n,j})$. It is convenient to recast the DFM (37)-(38) in matrix form as follows:

$$Y = \bar{X}\Phi_1 + \varepsilon \quad (39)$$

$$F = \tilde{F}\Phi_0 + \varepsilon_0 \quad (40)$$

where we define the $T \times n$ matrix $Y = (y_1, \dots, y_T)'$, the $T \times (k+1)$ matrix $\bar{X} = [\mathbf{1}'_T, F]$, with $\mathbf{1}_T$ being a $1 \times T$ vector of ones and $F = (f_1, \dots, f_T)'$ is a $T \times k$ matrix of factors, the $(k+1) \times n$ matrix $\Phi_1 = [a, \lambda^*]'$ and the $T \times n$ matrix of residuals is denoted as $\varepsilon = (\xi_1, \dots, \xi_T)'$, where $\varepsilon \sim \mathcal{N}(0, \Sigma_1)$. We define the $T \times kq$ matrix $\tilde{F} = (\tilde{f}_1, \dots, \tilde{f}_T)'$ with the $kq \times 1$ vectors $\tilde{f}_t = (f'_{t-1}, \dots, f'_{t-q})'$, the $kq \times k$ matrix $\Phi_0 = [\Phi_{0,1}, \dots, \Phi_{0,q}]'$, and the $T \times k$ matrix $\varepsilon_0 = (u_{0,1}, \dots, u_{0,T})'$.

Let us partition the parameter space Θ as $\theta_1 = \Phi_1$, $\theta_2 = \Sigma_1$, $\theta_3 = \Phi_0$, and $\theta_4 = \Sigma_0^*$. The prior for the constant terms and the factor loadings Φ_1 is usually selected to be normal, while

the prior for the Σ_1 is chosen to be an Inverted-Wishart. Furthermore, the priors for the parameters of the factor model (40) (i.e., Φ_0 and Σ_0) are chosen to belong to the $\mathcal{MN}\mathcal{TW}$ family. See for instance Otrok and Whiteman (1998).

Conditional on the factors, F , the system in (39) boils down to a multivariate linear Gaussian regression model. Hence, it is simple to see that the posterior density $(\Phi_1, \Sigma_1) | (\Phi_0, \Sigma_0^*, F, Y) = (\Phi_1, \Sigma_1) | (F, Y)$ belongs to the $\mathcal{MN}\mathcal{TW}$ family. Note that conditional on the factors, F , the likelihood function, $p(Y|\Theta)$ is not affected by the parameters Φ_0 and Σ_0^* , that is, $p(Y|\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1, F) = p(Y|\Phi_1, \Sigma_1, F)$. Therefore the posterior densities, $\Phi_0 | (\Phi_1, \Sigma_1, F, Y) = \Phi_0 | \Sigma_0$ and $\Sigma_0^* | (\Phi_1, \Sigma_1, F, Y) = \Sigma_0^*$, equal their priors and hence are analytically tractable.²⁰ Hence it follows that the *analytical tractability condition* is satisfied for $\tau = 4$.

In order to have that the *sampling condition* is satisfied, we need to show that it is possible to draw from the conditional posterior of factors $F | (\Phi_1, \Sigma_1, \Phi_0, \Sigma_0^*, Y)$. As discussed in Del Negro and Schorfheide (2010), one can draw from this density by using a variant of the Carter and Kohn (1994) approach applied to the state-space model (39)-(40) which is described in detail in Pitt and Kohn (2010).²¹

The application of the Chib's method to DFMs is straightforward. This method requires performing three reduced-Gibbs steps in addition to the Gibbs sampler. Method 1 follows equation (28). The likelihood $p(Y|\tilde{\Theta})$ does not have a closed-form solution but it can be easily evaluated through the Kalman filter. Method 1 approximates the joint posterior density $\hat{p}(\tilde{\Theta}|Y)$ from the output of Gibbs sampler as follows:

$$\hat{p}(\tilde{\Theta}|Y) = \frac{1}{n_r} \sum_{s=1}^{n_r} p(\tilde{\Phi}_1, \tilde{\Sigma}_1 | F^{(s)}, Y) p(\tilde{\Phi}_0, \tilde{\Sigma}_0 | F^{(s)}) \quad (41)$$

where the two densities on the right-hand side are known MNIW densities.

Method 2 computes:

$$\hat{p}_{M2}(Y) = \left[\frac{1}{n_r} \sum_{s=1}^{n_r} \frac{p(\tilde{\Theta} | F^{(s)}, Y)}{p(Y|\tilde{\Theta}) p(\tilde{\Theta})} \right]^{-1}$$

²⁰See Appendix F.

²¹See also Otrok and Whiteman (1998) for an alternative way of obtaining draws from the posterior distribution of factors. These scholars, first, derive an analytical expression for the joint Normal distribution of the observation Y and the factors, F : $p(Y, F | \bar{\Phi}, \Sigma, \Phi_0, \Sigma_0)$. Then, they use the formula for conditional means and covariance matrices to obtain the analytical expression for the conditional posterior distribution $F | \bar{\Phi}, \Sigma, \Phi_0, \Sigma_0, Y$.

where the draws $F^{(s)}$ are obtained from the Gibbs sampler. All the densities on the right-hand side have a known analytical characterization, except for the likelihood, which can be evaluated through the Kalman filter.

3.6 Factor-Augmented Vector Autoregressive Models (FAVARs)

Bernanke, Boivin, and Elias (2005) propose a hybrid model between a standard structural VAR model and a DFM model that has been called Factor-Augmented Vector Autoregression (henceforth FAVAR). This extension of the DFM paradigm allows for additional observations f_t^y in the measurement equation (37) such that

$$y_{i,t} = a_i + \lambda_i^y f_t^y + \lambda_i^f f_t^c + \xi_{i,t}, \quad t = 1, \dots, T \quad (42)$$

where λ_i^y is a $1 \times m$ vector and f_t^y is an $m \times 1$ vector, where f_t^c are the unobserved factors. For example, f_t^y might include the federal funds rate (as in Bernanke, Boivin and Elias, 2005) or other policy instruments, such as monetary aggregates (as in Ahmadi and Ritschl, 2009).

The joint dynamics of the perfectly observable vector, f_t^y , and the unobserved factors, f_t^c , are described by the following state equation

$$\begin{bmatrix} f_t^c \\ f_t^y \end{bmatrix} = \Phi_{0,1} \begin{bmatrix} f_{t-1}^c \\ f_{t-1}^y \end{bmatrix} + \dots + \Phi_{0,q} \begin{bmatrix} f_{t-q}^c \\ f_{t-q}^y \end{bmatrix} + u_{0,t}, \quad u_{0,t} \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma_0) \quad (43)$$

which is a VAR(q) for (f_t^c, f_t^y) . Comparing equation (43) with equation (38), we have that now the matrices $\Phi_{0,j}$ have dimension $(k+m) \times (k+m)$.

The model described by equations (42)-(43) is unidentified and cannot be estimated. Identification is achieved by imposing restrictions on factors and their coefficients in equation (42). In particular, to avoid indeterminacy of the model, Bernanke, Boivin, and Elias (2005) impose the following normalization scheme: (i) the upper $k \times k$ block of the $n \times k$ matrix of factor loadings is an identity matrix and (ii) the upper $k \times m$ block of the $n \times m$ matrix of coefficients for the observed vector is composed of zeros.

Let us recast equation (42) as follows:

$$X_t = A + \lambda f_t + e_t \quad e_t \stackrel{iid}{\sim} \mathcal{N}(0, R)$$

where $A = (a_1, \dots, a_n, \mathbf{0}_{m \times 1})'$, $X_t = (y_t', f_t^{y'})'$, $f_t = (f_t^c, f_t^{y'})'$, $e_t = (\xi_{1,t}, \dots, \xi_{n,t}, \mathbf{0}_{m \times 1})'$,

$$R = \begin{bmatrix} \Sigma & \mathbf{0}_{n \times (m-n)} \\ \mathbf{0}_{(m-n) \times n} & \mathbf{0}_{(m-n) \times (m-n)} \end{bmatrix}, \quad \Sigma \text{ is an } n \times n \text{ matrix}$$

and

$$\lambda = \begin{bmatrix} \lambda^f & \lambda^y \\ \mathbf{0}_{m \times k} & I_m \end{bmatrix}$$

with $\lambda^f = (\lambda_1^f, \dots, \lambda_k^f)'$ and $\lambda^y = (\lambda_1^y, \dots, \lambda_m^y)'$. Therefore, λ^f is an $n \times k$ matrix and λ^y is an $n \times m$ matrix.

Given equations (42)-(43), we can recast the FAVAR model in matrix form as we did for DFM models in equations (39)-(40) where now the matrix of observables is given by the $T \times (n + m)$ matrix X instead of just the $T \times n$ matrix Y . Therefore, it directly follows that both the *sampling* and *analytical tractability conditions* are satisfied. Application of the estimators discussed in the paper is straightforward.

4 Empirical Application

In this section, we assess the gains in accuracy and computational burden of the two methods proposed in the paper by means of fitting the VAR model in (12)-(13) with $l = 1$ (i.e., linear time trend model) to macroeconomic times series. We focus on this type of models because the true conditional predictive density $Y|\tilde{\Gamma}$, defined in (18), is known in closed-form. Therefore, comparing the true conditional predictive density with the one estimated by Chib's method provides a natural way to shed light on the gains linked to our estimators.

In particular, we are interested in assessing the gains in accuracy and computational burden as the dimensionality of the model varies. To do so, we fit four VAR models to six encompassing data sets. In particular, we fit autoregressive models with lags $p = 1, \dots, 4$ to data sets containing from one up to six variables. Let us enumerate all series under analysis: Real Gross Domestic Product (source: Bureau of Economic Analysis, *GDPC96*), Implicit Price Deflator (source: Bureau of Economic Analysis, *GDPDEF*), Personal Consumption Expenditures (source: Bureau of Economic Analysis, *PCEC*), Fixed Private Investment (source: Bureau of Economic Analysis, *FPI*), Effective Federal Funds Rate (source: Board of Governors of the Federal Reserve System, *FEDFUNDS*), and Average Weekly Hours Duration in the Non-farm Business (source: U.S. Department of Labor, *PRS85006023*). All

data are quarterly²². The data set ranges from 1954:1 to 2008:4. Table 1 describes the data series contained in each data set.

We elicit the prior density for the parameters of the VAR in deviations, (Φ, Σ) , by using the single-unit-root prior, suggested by Sims and Zha (1998). To pin down this prior, we need to choose the value of five hyperparameters²³. We follow Del Negro and Schorfheide (2004), Giannone, Lenza, and Primiceri (2010), and Carriero, Kapetanios, and Marcellino (2010) setting these hyperparameters so as to maximize the conditional predictive density, $p(Y|\tilde{\Gamma})$. To this end, we perform a stochastic search based on simulated annealing (Judd, 1998) with 1,000 stochastic draws²⁴. Furthermore, the prior density depends on the first and second moments of some pre-sample data. We use the moments of a pre-sample ranging from 1947:1 to 1953:4.

We run ten chains of m number of draws in the Gibbs sampler and in the reduced-Gibbs sampler, where $m = \{100, 1,000, 10,000, 100,000\}$. We also run one chain with one million draws.

4.1 Gains in Accuracy from Method 1

Our estimators rely on the insight that exploiting the *analytical tractability* condition increases the accuracy of MDD estimators. In this empirical application, we assess the inaccuracy associated with neglecting the *analytical tractability condition*. Consider the VAR model of the form (12)-(13). In this framework, Method 1 differs from Chib’s method only on the computation of the conditional predictive density, $p(Y|\tilde{\Gamma})$ defined in (18). The *analytical tractability condition* implies that this predictive density has a known analytical characterization. Method 1 exactly calculates the conditional predictive density $p(Y|\tilde{\Gamma})$ via its analytical expression, given in equation (18). Chib’s method, conversely, approximates such conditional predictive density computationally via equation (19) that requires performing the reduced Gibbs step described in Algorithm 3. Thus, the inaccuracy deriving from neglecting the *analytical tractability condition* can be quantified by the gap between the estimated conditional predictive density using Chib’s approach, $\hat{p}_{\text{CHIB}}(Y|\tilde{\Gamma})$, and its true value,

²²Data on the Effective Federal Funds Rate are obtained as average of daily figures.

²³The first hyperparameter sets the overall tightness of the prior. The second hyperparameter controls the variance for the coefficients of the lagged variables. The third hyperparameter establishes the weight for the prior for the variance and covariance matrix of residuals. Finally, the other two hyperparameters affect the persistence of the prior-dummy observations. See Del Negro and Schorfheide (2010), Section 2.2, for more details.

²⁴During the search, we apply Method 1 to evaluate the analytical part of the MDD.

$p(Y|\tilde{\Gamma})$. Note that, as the number of draws in the reduced Gibbs step, n_r , goes to infinity, the size of the gap goes to zero, that is, $\lim_{n_r \rightarrow \infty} \hat{p}_{CHIB}(Y|\tilde{\Gamma}) = p(Y|\tilde{\Gamma})$. In this application, we assess the convergence of Chib’s method to the true conditional predictive density by computing

$$\left| \log \left(\hat{p}_{CHIB}(Y|\tilde{\Gamma}) \right) - \log \left(p(Y|\tilde{\Gamma}) \right) \right| \quad (44)$$

We refer to this difference as the estimation bias for the conditional predictive density.

We set the value of the parameter block $\tilde{\Gamma}$ to be equal to the OLS estimator²⁵. Given this restriction, we perform the reduced Gibbs step and compute the conditional predictive density $\hat{p}_{CHIB}(Y|\tilde{\Gamma})$. We compute the absolute difference in (44) for every chain, VAR model ($p = 1, \dots, 4$), and data set.

Figure 1 reports the (across-chain mean of the) estimation bias for the conditional predictive density for the 24 models of interest when performing 1,000,000 draws in both the Gibbs sampler and the reduced Gibbs step²⁶. We find worth emphasizing the following two results. First, for a given number of lags p , the estimation bias grows at an increasing rate as the number of observable variables increases. Second, for a given number of observables, the estimation bias grows at an increasing rate as the number of lags p increases. For example, the size of the gap for a six-variate VAR(4) is about 9 times the size of the bias for the VAR(1) model.

We document in Table 2 how the estimation bias varies as one increases the number of draws in the reduced-Gibbs step performed by Chib’s method for six-variate VARs models. We conclude that for a given data set and a given model, the bias is quite stable despite the increase in the number of posterior draws in the reduced-Gibbs step. This suggests that the MC integration in (16) exhibits a rather slow convergence.

4.2 Model Selection

In this section, we turn our attention to the crucial issue of the effect of inaccurate estimates when performing Bayesian model selection. Given a loss function that reflects the preferences of the econometrician and a set of candidate models, the optimal decision is to select the model that minimizes the posterior expected loss function (Schorfheide, 2000). Under a 0-1 loss function, selecting the model with the largest posterior probability can be easily shown

²⁵This restriction will be relaxed in the experiment conducted in the next section.

²⁶Results for $n = \{100; 1,000; 10,000; 100,000; 1,000,000\}$, where n is the number of draws, are available upon request.

to be the optimal decision. Let us define the model set to be formed by the four VAR models, that is, $\{VAR(p), 1 \leq p \leq 4\}$ ²⁷. Furthermore, we assume that the prior model probabilities, $\{\pi_{p,0}, 1 \leq p \leq 4\}$, are the same across the four candidate models. The posterior probability of the VAR(p), with $0 \leq p \leq 4$, \mathcal{M}_p , is given by:

$$\pi_{p,T} = \frac{\pi_{p,0} \cdot p(Y|\mathcal{M}_p)}{\sum_{i=1}^4 \pi_{i,0} \cdot p(Y|\mathcal{M}_i)} \quad (45)$$

where $\pi_{p,T}$ ($\pi_{p,0}$) stands for the posterior (prior) probability of the VAR(p) and $p(Y|\mathcal{M}_p)$ denotes the MDD of the VAR(p).

For every estimator, we permute MDDs estimated at each chain across the four VAR models which delivers 10,000 quadruplets of posterior probabilities computed using (45). Figure 2 reports the distributions for these 10,000 posterior probabilities computed by the three estimators. The distributions of the posterior probabilities associated with the VAR(1) and the VAR(2) are a mass point at zero, suggesting that all methods strongly disfavor the VAR(1) and the VAR(2). Furthermore, while both Method 1 and Method 2 strongly favor the VAR(4), the distribution related to Chib’s method peaks at 20%. Conversely, Chib’s method strongly favors the VAR(3) model with a median posterior probability of about 80%. This shows that the estimation bias due to a fully computational approach may significantly distort model rankings.

Two important remarks about Figure 2 are in order. First, since Method 1 and Chib’s estimator differ only in how they calculate the conditional posterior $\Sigma|\Gamma, Y$, the bias in model ranking must be due to the inaccuracy in the MC integration (16), based on the reduced Gibbs step. Second, although Method 1 and 2 estimate the MDD through different approaches²⁸, these two methods deliver posterior model rankings that are remarkably similar. Hence, the accuracy of the two methods proposed in the paper is of the same order of magnitude.

Let us analyze the stability of the three estimators under analysis. Tables 3-6 report the across-chain means and standard deviations of the log MDD for each of estimators, models, and data sets. We can conclude that at 10,000 draws, the stability of all the three estimators is quite good already. This result suggests that increasing the number of draws is unlikely to

²⁷We have extended the exercise to include VAR(5) and VAR(6). The results of this extended exercise are available upon request. We have decided to not present them in the paper because all the three estimators deliver very small MDDs for these two models. Hence, all the results discussed in this section are unchanged.

²⁸Recall that Method 1 exploits the fact that the MDD can be expressed as the normalizing constant of the joint posterior density for model parameters. In contrast, Method 2 relies on the principle of reciprocal importance sampling.

change the predictions about which model attains the largest posterior probability. In other words, no sizeable corrections have to be expected from increasing the number of draws in the reduced-Gibbs sampler. This last finding is in line with the slow convergence of the MC integration based on the reduced-Gibbs step draws, discussed in Section 4.1.

4.3 Computation Time

Figures 3-5 show how the computation time (in seconds) associated with the three estimators under analysis varies as the number of observable variables and the number of lags, p increases. Comparing these figures, we observe that Method 2 is computationally more convenient than Method 1 and Chib’s method for any model specification and any data set. In Figure 5, we observe that for Method 2 (i) the computing time is almost invariant to the number of lags included in the model and (ii) the increases in computing time due to the inclusion of additional observable variables are quite small. Quite remarkably, estimating the MDD associated with a six-variate VAR(4) with the Method 2 and 100,000 posterior draws,²⁹ takes less than 1/10 seconds. This result is striking but not surprising since Method 2 only requires performing the Gibbs Sampler regardless the number of partition that can be integrated out.

Furthermore, if one compares Figures 3-5, one would note that the computing time associated with our estimators (i.e., Method 1 and Method 2) is not growing exponentially as one increases the number of observables or the number of lags of the VAR. Moreover, we report in Figure 6 the difference in computing time between Chib’s method and Method 1. Recall that these two estimators only differ in how they calculate the conditional posterior $\Sigma|\Gamma, Y$. Hence, Figure 6 shows how the computing time to perform the reduced-Gibbs step changes as the number of lags or observables in the VAR model varies. One can see that the reduced-Gibbs step is the culprit for the computing time associated with the Chib’s method to grow exponentially with respect to the number of observables and lags. In contrast, it follows that exploiting the *analytical tractability condition* prevents our two estimators from being affected by such a curse of dimensionality.

²⁹100,000 draws ensure very reliable estimates as the small size of the across-chain standard deviation for the MC experiment, reported in Table 6, suggests.

5 Concluding Remarks

The paper develops two new estimators for the marginal likelihood of the data. These estimators are shown to apply to a broad set of popular models in time series econometrics: Vector AutoRegressive Models (VARs), Reduced Rank Regression Models such as Vector Equilibrium Correction Models (VECMs), Markov-Switching VAR models (MS VARs), Time-Varying Parameter VAR models (TVP VARs), Dynamic Factor Models (DFMs), and Factor Augmented VAR models (FAVARs). Our estimators rely on the fact that it is possible to analytically integrate out one or more parameter blocks from the block-conditional posterior densities implied by those models.

An empirical application based on a standard macro data set reveals that our estimators translate into significant gains in accuracy and computational burden when compared to a very popular fully-computational approach. We find that the estimation bias associated with the fully-computational estimator may severely distort model rankings. Furthermore, our estimators do not suffer the curse of dimensionality that affects the fully-computational method. In particular, Method 2 is fast enough to be well-suited for applications where the marginal likelihood of VAR models has to be computed several times (e.g., Bayesian selection or average across a large set of models).

To sum up, the paper favors the idea that estimators that are tailored to the specific features of a model at hand are likely to dominate universal estimators, which are virtually applicable to a broader set of models but have to rely on brute-force computational methods. Using estimators that exploit the specific features of a model at hand is very rewarding, especially when the models in question are quite densely parameterized

References

- AHMADI, P. A., AND A. RITSCHL (2009): “Depression Econometrics: A FAVAR Model of Monetary Policy During the Great Depression,” SFB 649 Discussion Paper 2009-054.
- ARUOBA, S. B., F. X. DIEBOLD, AND C. SCOTTI (2009): “Real-Time Measurement of Business Conditions,” *Journal of Business and Economic Statistics*, 24(4), 417–427.
- AYHAN, K. M., C. OTROK, AND C. WHITEMAN (2003): “International Business Cycles: World, Region and Country Specific Factors,” *American Economic Review*, 93(4), 1216–1239.
- (2008): “Understanding the Evolution of World Business Cycles,” *Journal of International Economics*, 75, 110–130.
- BANBURA, M., D. GIANNONE, AND L. REICHLIN (2010): “Large Bayesian Vector Auto Regressions,” *Journal of Applied Econometrics*, 25(1), 71–92.
- BERNANKE, B. S., J. BOIVIN, AND P. ELIASZ (2005): “Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach,” *Quarterly Journal of Economics*, 120(1), 387–422.
- BOIVIN, J., AND S. NG (2005): “Understanding and Comparing Factor Based Macroeconomic Forecasts,” *International Journal of Central Banking*, 1, 117–152.
- CARRIERO, A., G. KAPETANIOS, AND M. MARCELLINO (2010): “Forecasting Government Bond Yields with Large Bayesian VARs,” CEPR Discussion Paper No. 7796.
- CARTER, C., AND R. KOHN (1994): “On Gibbs Sampling for State Space Models,” *Biometrika*, 81(3), 541–553.
- CHIB, S. (1995): “Marginal Likelihood from the Gibbs Output,” *Journal of the American Statistical Association*, 90(432), 1313–1321.
- COGLEY, T., AND T. J. SARGENT (2002): “Evolving post World War II US inflation dynamics,” in *NBER Macroeconomics Annual 2001*, ed. by B. S. Bernanke, and K. Rogoff, vol. 16, pp. 331–388. MIT Press, Cambridge.
- (2005): “Drifts and volatilities: monetary policies and outcomes in the post WWII US,” *Review of Economic Dynamics*, 8, 262–302.

- DEL NEGRO, M., AND F. SCHORFHEIDE (2004): “Priors from General Equilibrium Models for VARS,” *International Economic Review*, 45(2), 643–673.
- (2010): “Bayesian Macroeconometrics,” in *The Handbook of Bayesian Econometrics*, ed. by H. K. van Dijk, J. F. Geweke, and G. Koop. Oxford University Press.
- DOAN, T., R. LITTERMAN, AND C. A. SIMS (1984): “Forecasting and Conditional Projection Using Realistic Prior Distributions,” *Econometric Reviews*, 3(1), 1–100.
- FIorentini, G., C. PLANAS, AND A. ROSSI (2011): “The marginal likelihood of dynamic mixture models: some new results,” Mimeo.
- Forni, M., M. Hallin, M. Lippi, AND L. Reichlin (2003): “Do Financial variables help forecasting inflation and real activity in the Euro Area?,” *Journal of Monetary Economics*, 50, 1243–1255.
- GELFAND, A. E., AND D. K. DEY (1994): “Bayesian Model Choice: Asymptotics and Exact Calculations,” *Journal of the Royal Statistical Society B*, 56, 501–514.
- GELFAND, A. E., A. F. M. SMITH, AND T.-M. LEE (1992): “Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling,” *Journal of the American Statistical Association*, 87(418), 523–532.
- GEWEKE, J. (1977): “The Dynamic Factor Analysis of Economic Time Series,” in *Latent Variables in Socio-Economic Models*, ed. by D. J. Aigner, and A. S. Goldberg. Chap. 19, North Holland, Amsterdam.
- (1996): “Bayesian Reduced Rank Regression in Econometrics,” *Journal of Econometrics*, 75(1), 121–146.
- GEWEKE, J., AND G. ZHOU (1996): “Measuring the Pricing Error of the Arbitrage Pricing Theory,” *Review of Financial Studies*, 9(2), 557–587.
- GEWEKE, J. F. (1999): “Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication,” *Econometric Reviews*, 18, 1–126.
- GIANNONE, D., M. LENZA, AND G. PRIMICERI (2010): “Prior Selection for Vector Autoregressions,” mimeo.
- GIANNONE, D., L. REICHLIN, AND D. SMALL (2008): “Nowcasting: The Real-Time Informational Content of Macroeconomic Data,” *Journal of Monetary Economics*, 55, 665–676.

- HAMILTON, J. D. (1989): “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle,” *Econometrica*, 57(2), 357–384.
- JUDD, K. L. (1998): *Numerical Methods in Economics*. The MIT Press, Boston.
- KIM, C.-J., AND C. R. NELSON (1999): *State Space Models with Regime Switching*. MIT Press, Cambridge, MA.
- KOOP, G. (2011): “Forecasting with Medium and Large Bayesian VARs,” mimeo University of Strathclyde.
- KOOP, G., AND S. POTTER (2004): “Forecasting in Dynamic Factor Models Using Bayesian Model Averaging,” *Econometric Journal*, 7(2), 550–565.
- KOOP, G., R. STRACHAN, H. VAN DIJK, AND M. VILLANI (2006): “Bayesian Approaches to Cointegration,” in *Palgrave Handbook of Econometrics*, ed. by T. C. Mills, and K. P. Patterson, vol. 1, pp. 871–898. Palgrave Macmillan, Basingstoke, United Kingdom.
- KOROBILIS, D. (forthcoming): “Forecasting in Vector Autoregressions with many predictors,” *Advances in Econometrics, Vol 23: Bayesian Macroeconometrics*.
- LITTERMAN, R. B. (1980): “Techniques for Forecasting with Vector Autoregressions,” Ph.D. thesis, University of Minnesota.
- MENG, X.-L., AND W. H. WONG (1996): “Simulating Ratios of Normalizing Constants Via a Simple Identity: A Theoretical Exploration,” *Statistica Sinica*, 6, 831–860.
- NEWTON, M. A., AND A. E. RAFTERY (1999): “Approximate Bayesian Inference by the Weighted Likelihood Bootstrap,” *Journal of the Royal Statistical Society B*, 56(1), 3–48.
- OTROK, C., AND C. H. WHITEMAN (1998): “Bayesian Leading Indicators: Measuring and Predicting Economic Conditions in Iowa,” *International Economic Review*, 39(4), 997–1014.
- PITT, P. G. M. K., AND R. KOHN (2010): “Bayesian Inference for Time Series State Space Models,” in *The Handbook of Bayesian Econometrics*, ed. by H. K. van Dijk, J. F. Geweke, and G. Koop. Oxford University Press.
- PRIMICERI, G. (2005): “Time Varying Structural Vector Autoregressions and Monetary Policy,” *Review of Economic Studies*, 72(3), 821–852.

- SARGENT, T. J., AND C. A. SIMS (1977): “Business Cycle Modeling Without Pretending to Have Too Much a Priori Economic Theory,” in *New Methods in Business Cycle Research*. FRB Minneapolis, Minneapolis.
- SCHORFHEIDE, F. (2000): “Loss Function-Based Evaluation of DSGE Models,” *Journal of Applied Econometrics*, 15(6), 645–670.
- SIMS, C. A. (1980): “Macroeconomics and Reality,” *Econometrica*, 48(4), 1–48.
- SIMS, C. A., AND T. ZHA (1998): “Bayesian Methods For Dynamic Multivariate Models,” *International Economic Review*, 39(4), 949–968.
- STOCK, J., AND M. WATSON (1999): “Forecasting Inflation,” *Journal of Monetary Economics*, 44, 293–335.
- (2002a): “Forecasting Using Principal Components for a Large Number of Predictors,” *Journal of the American Statistical Association*, 97, 1167–1179.
- STOCK, J., AND M. WATSON (2002b): “Macroeconomic Forecasting Using Diffusion Indexes,” *Journal of Business and Economic Statistics*, 20, 147–162.
- STOCK, J., AND M. WATSON (2002c): “Macroeconomic Forecasting Using Diffusion Indexes,” *Journal of Business and Economic Statistics*, 20, 147–162.
- STOCK, J. H., AND M. W. WATSON (1989): “New Indices of Coincident and Leading Economic Estimators,” in *NBER Macroeconomics Annual 1989*, ed. by O. J. Blanchard, and S. Fisher, vol. 4, pp. 351–394. MIT Press, Cambridge.
- VILLANI, M. (2009): “Steady State Priors for Vector Autoregressions,” *Journal of Applied Econometrics*, 24(4), 630–650.

Figures and Tables

Figure 1: ESTIMATION BIAS FOR THE CONDITIONAL PREDICTIVE DENSITY (1,000,000 DRAW)

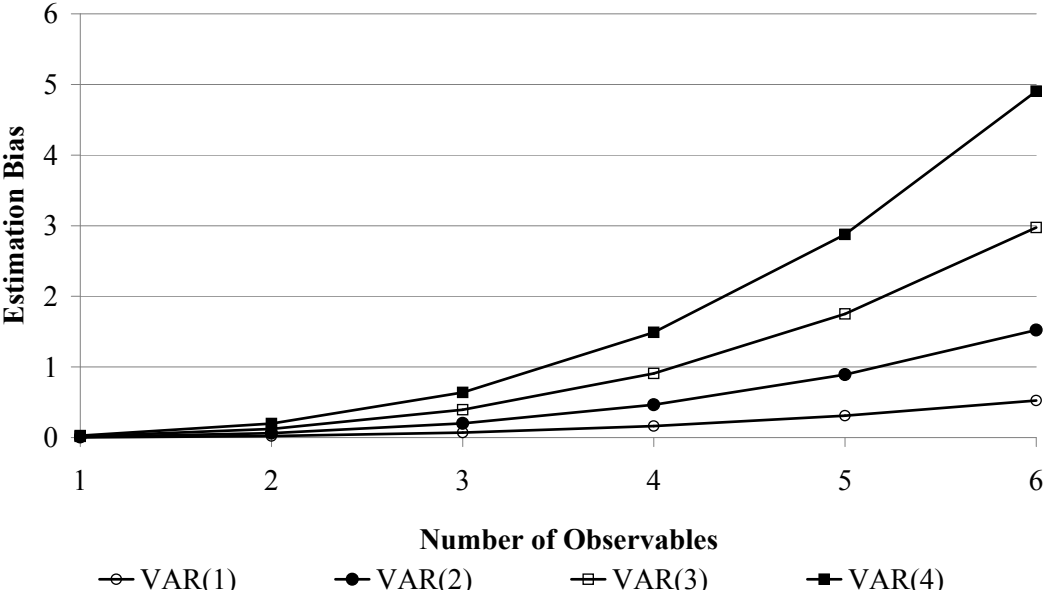
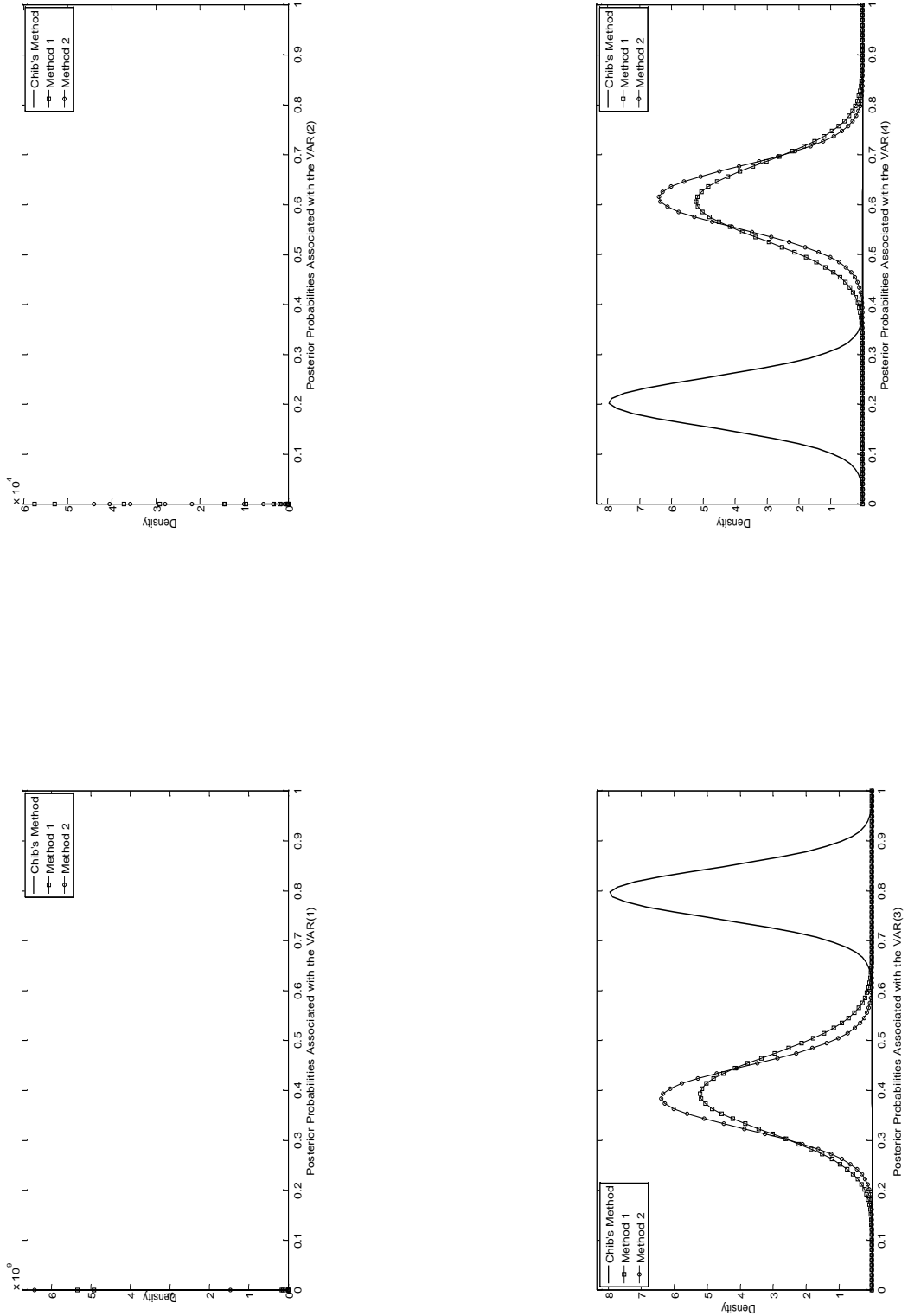
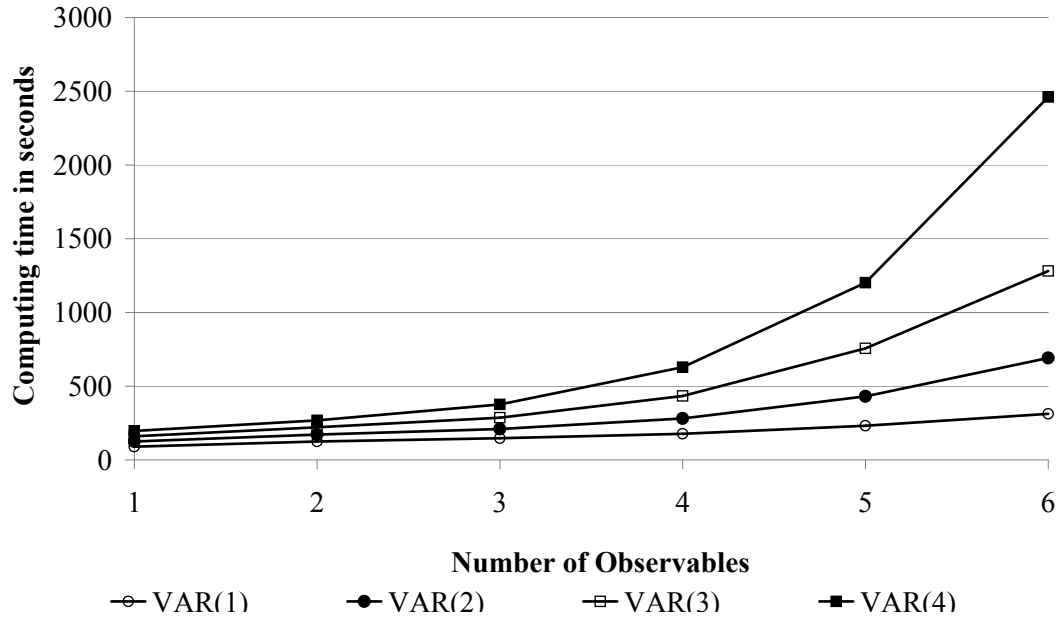


Figure 2: DISTRIBUTION OF POSTERIOR PROBABILITIES FOR VAR(P), P=(1,2,3,4).



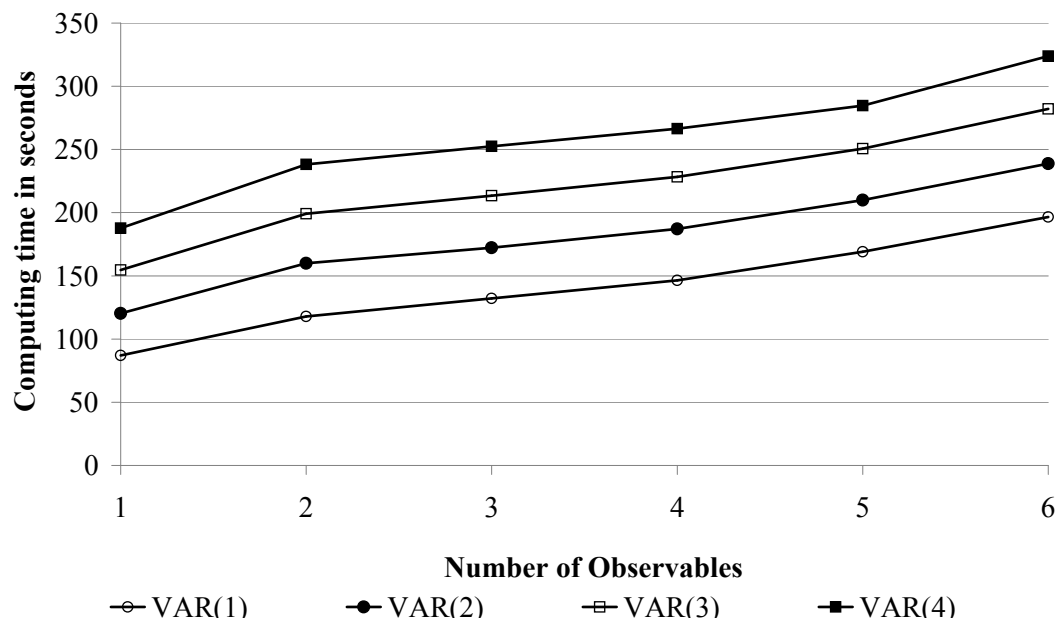
Notes: The distributions are obtained by fitting a Gaussian distribution to the 10,000 posterior probabilities delivered by the Monte Carlo experiment, described in Section 4.2.

Figure 3: COMPUTING TIME (IN SECONDS): CHIB'S METHOD



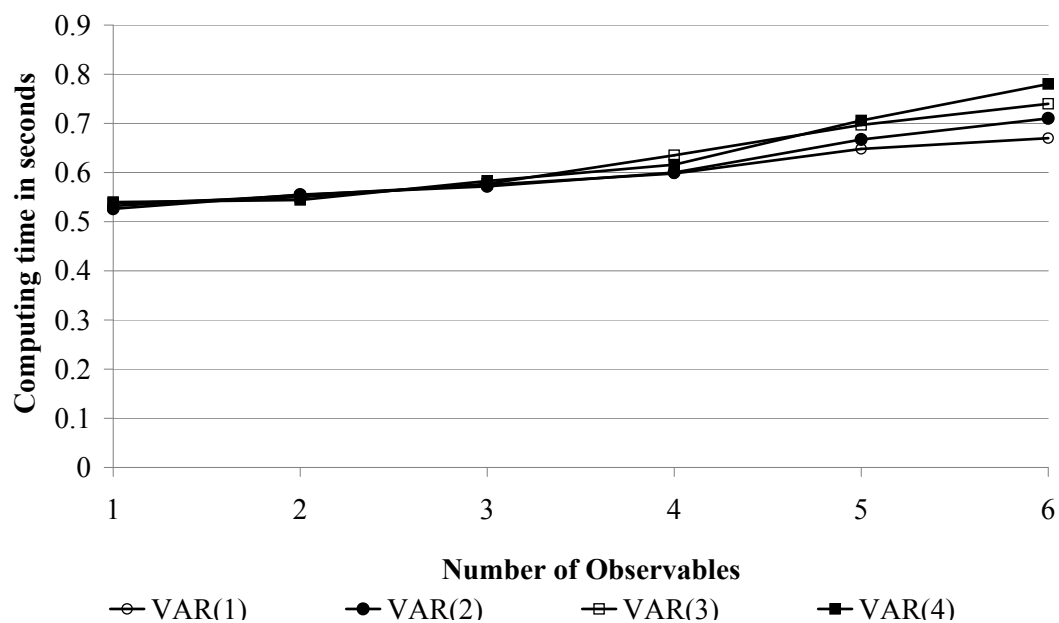
Notes: We use 100,000 draws in both the Gibbs sampler and the reduced-Gibbs step

Figure 4: COMPUTING TIME (IN SECONDS): METHOD 1



Notes: We use 100,000 draws in both the Gibbs sampler

Figure 5: COMPUTING TIME (IN SECONDS): METHOD 2



Notes: We use 100,000 draws in both the Gibbs sampler

Figure 6: ACROSS-CHAIN AVERAGE OF COMPUTING TIME (IN SECONDS) FOR CHIB'S METHOD RELATIVE TO METHOD 1

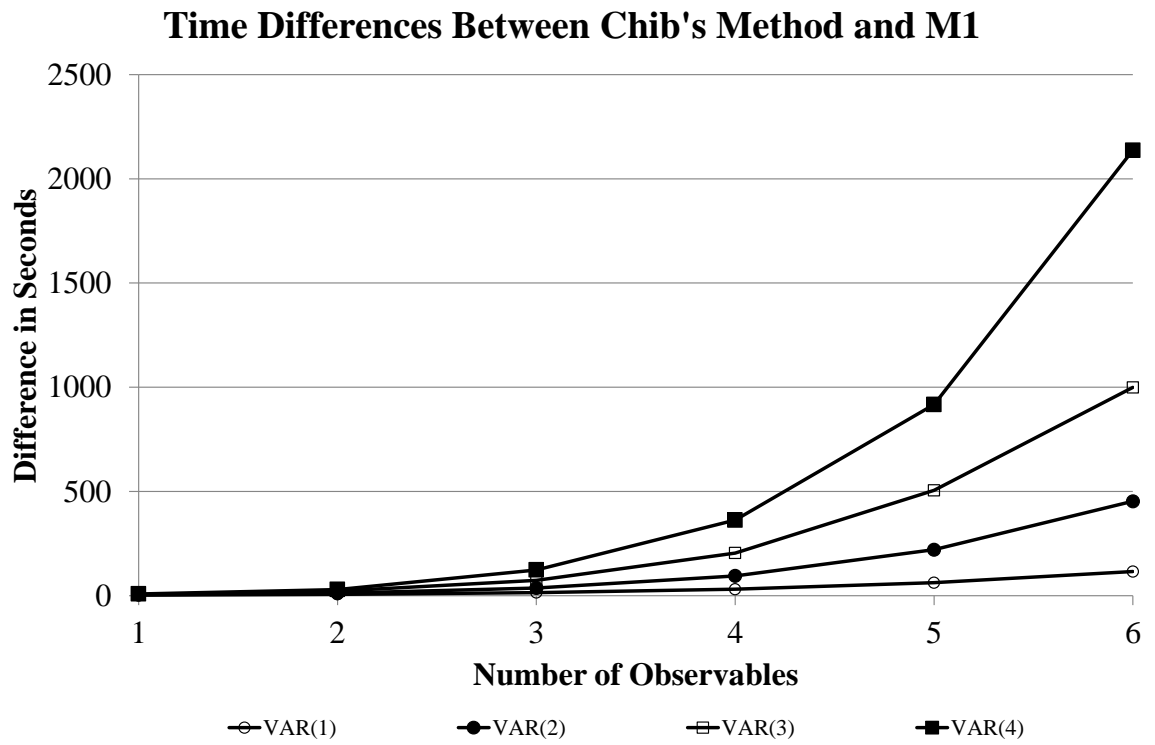


Table 1: MODELS

| Series | One-variate | Bi-variate | Three-variate | Four-variate | Five-variate | Six-variate |
|-----------------------------------|-------------|------------|---------------|--------------|--------------|-------------|
| Real Gross Domestic Product | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Implicit Price Deflator | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Personal Consumption Expenditures | | | ✓ | ✓ | ✓ | ✓ |
| Fixed Private Investment | | | | ✓ | ✓ | ✓ |
| Effective Federal Funds Rate | | | | | ✓ | ✓ |
| Average Weekly Hours | | | | | | ✓ |

Table 2: ACROSS-CHAIN AVERAGES OF THE ESTIMATION BIAS FOR THE CONDITIONAL PREDICTIVE DENSITY: SIX-VARIATE VAR

| Draws | VAR(1) | VAR(2) | VAR(3) | VAR(4) |
|-------------|--------------|--------------|--------------|--------------|
| 100 | 0.53 | 1.48 | 2.69 | 4.49 |
| | <i>0.107</i> | <i>0.269</i> | <i>0.219</i> | <i>0.489</i> |
| 1, 000 | 0.55 | 1.52 | 2.91 | 4.53 |
| | <i>0.053</i> | <i>0.060</i> | <i>0.175</i> | <i>0.218</i> |
| 10, 000 | 0.53 | 1.53 | 2.97 | 4.90 |
| | <i>0.009</i> | <i>0.023</i> | <i>0.058</i> | <i>0.183</i> |
| 100, 000 | 0.53 | 1.52 | 2.99 | 4.91 |
| | <i>0.003</i> | <i>0.006</i> | <i>0.033</i> | <i>0.063</i> |
| 1, 000, 000 | 0.53 | 1.52 | 2.97 | 4.90 |

Notes: Across-chain means of absolute differences. Numerical standard errors in italics. Draws refers to the number of posterior draws and the number of draws in the reduced Gibbs step. For one million draws, we do not report numerical standard errors.

Table 3: LOG-MARGINAL DATA DENSITY

| Model | Draws | One-variate | | Two-variate | | Three-variate | | | | |
|--------|-----------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|----------------------|----------------------|----------------------|
| | | Chib | Method 1 | Method 2 | Chib | Method 1 | Method 2 | Chib | Method 1 | Method 2 |
| VAR(1) | 10 ² | -381.477 (0.054) | -381.471 (0.055) | -381.568 (0.049) | -988.066 (0.084) | -988.039 (0.082) | -988.124 (0.059) | -1141.400 (0.113) | -1141.300 (0.101) | -1141.703 (0.117) |
| | 10 ³ | -381.489 (0.017) | -381.485 (0.017) | -381.493 (0.017) | -988.009 (0.019) | -988.039 (0.023) | -988.013 (0.019) | -1141.400 (0.022) | -1141.319 (0.024) | -1141.369 (0.022) |
| | 10 ⁴ | -381.489 (0.009) | -381.484 (0.009) | -381.490 (0.009) | -988.009 (0.010) | -987.983 (0.009) | -987.994 (0.010) | -1141.397 (0.008) | -1141.317 (0.004) | -1141.338 (0.007) |
| | 10 ⁵ | -381.488 (0.002) | -381.484 (0.002) | -381.488 (0.002) | -988.008 (0.004) | -987.982 (0.004) | -987.994 (0.004) | -1141.396 (0.002) | -1141.318 (0.002) | -1141.337 (0.002) |
| | 10 ⁶ | -381.487 | -381.483 | -381.489 | -988.007 | -987.981 | -987.994 | -1141.396 | -1141.318 | -1141.337 |
| | VAR(2) | 10 ² | -381.593 (0.040) | -381.590 (0.039) | -381.648 (0.078) | -988.369 (0.119) | -988.301 (0.092) | -988.400 (0.075) | -1142.127 (0.098) | -1141.900 (0.091) |
| | 10 ³ | -381.588 (0.023) | -381.579 (0.023) | -381.599 (0.021) | -988.342 (0.034) | -988.274 (0.031) | -988.319 (0.020) | -1142.065 (0.042) | -1141.849 (0.041) | -1141.902 (0.016) |
| | 10 ⁴ | -381.588 (0.007) | -381.577 (0.007) | -381.595 (0.003) | -988.340 (0.009) | -988.269 (0.008) | -988.299 (0.008) | -1142.054 (0.016) | -1141.841 (0.013) | -1141.885 (0.006) |
| | 10 ⁵ | -381.592 (0.003) | -381.581 (0.003) | -381.594 (0.002) | -988.339 (0.002) | -988.270 (0.002) | -988.297 (0.002) | -1142.055 (0.005) | -1141.841 (0.005) | -1141.881 (0.002) |
| | 10 ⁶ | -381.593 | -381.581 | -381.593 | -988.338 | -988.268 | -988.296 | -1142.054 | -1141.839 | -1141.882 |

Notes: Draws refers to the number of posterior draws and the number of draws in the reduced Gibbs step. Across-chain standard deviations are reported within brackets.

Table 4: LOG-MARGINAL DATA DENSITY

| Model | Draws | One-variate | | Two-variate | | Three-variate | | | | |
|--------|-----------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|----------------------|----------------------|----------------------|
| | | Chib | Method 1 | Method 2 | Chib | Method 1 | Method 2 | Chib | Method 1 | Method 2 |
| VAR(3) | 10 ² | -381.597 (0.119) | -381.580 (0.118) | -381.687 (0.007) | -988.457 (0.085) | -988.351 (0.102) | -988.580 (0.074) | -1142.390 (0.163) | -1142.007 (0.096) | -1142.420 (0.182) |
| | 10 ³ | -381.633 (0.034) | -381.615 (0.034) | -381.642 (0.021) | -988.525 (0.040) | -988.392 (0.037) | -988.426 (0.019) | -1142.444 (0.046) | -1142.045 (0.039) | -1142.137 (0.025) |
| | 10 ⁴ | -381.641 (0.009) | -381.621 (0.009) | -381.638 (0.003) | -988.508 (0.009) | -988.380 (0.006) | -988.426 (0.007) | -1142.452 (0.010) | -1142.042 (0.011) | -1142.107 (0.007) |
| | 10 ⁵ | -381.640 (0.003) | -381.620 (0.003) | -381.641 (0.002) | -988.513 (0.003) | -988.383 (0.002) | -988.424 (0.002) | -1142.449 (0.006) | -1142.038 (0.004) | -1142.107 (0.002) |
| | 10 ⁶ | -381.639 | -381.619 | -381.640 | -988.512 | -988.381 | -988.425 | -1142.450 | -1142.040 | -1142.106 |
| | VAR(4) | 10 ² | -381.607 (0.132) | -381.590 (0.111) | -381.626 (0.060) | -988.531 (0.105) | -988.289 (0.065) | -988.497 (0.161) | -1142.502 (0.230) | -1141.917 (0.121) |
| | 10 ³ | -381.600 (0.024) | -381.572 (0.019) | -381.611 (0.016) | -988.477 (0.018) | -988.278 (0.022) | -988.363 (0.017) | -1142.544 (0.063) | -1141.871 (0.039) | -1141.985 (0.028) |
| | 10 ⁴ | -381.619 (0.004) | -381.587 (0.005) | -381.609 (0.005) | -988.483 (0.015) | -988.281 (0.009) | -988.341 (0.007) | -1142.526 (0.019) | -1141.865 (0.011) | 1141.956 (0.008) |
| | 10 ⁵ | -381.614 (0.003) | -381.582 (0.003) | -381.610 (0.001) | -988.494 (0.003) | -988.282 (0.003) | -988.339 (0.002) | -1142.532 (0.003) | -1141.868 (0.002) | -1141.954 (0.002) |
| | 10 ⁶ | -381.612 | -381.581 | -381.610 | -988.495 | -988.284 | -988.338 | -1142.534 | -1141.868 | -1141.954 |

Notes: Draws refers to the number of posterior draws and the number of draws in the reduced Gibbs step. Across-chain standard deviations are reported within brackets.

Table 5: LOG-MARGINAL DATA DENSITY

| Model | Draws | Four-variate | | Five-variate | | Six-variate | | | | |
|--------|-----------------|--------------------|--------------------|--------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | | Chib | Method 1 | Method 2 | Chib | Method 1 | Method 2 | Chib | Method 1 | Method 2 |
| VAR(1) | 10 ² | 571.197 (0.384) | 571.339 (0.372) | 571.180 (0.273) | 1333.685 (1.520) | 1334.016 (1.577) | 1332.657 (1.481) | 3660.218 (6.996) | 3661.297 (8.021) | 3652.205 (2.883) |
| | 10 ³ | 571.236 (0.132) | 571.393 (0.129) | 571.688 (0.131) | 1334.535 (0.420) | 1334.746 (0.423) | 1335.038 (0.420) | 3653.757 (1.905) | 3654.158 (1.890) | 3655.653 (1.905) |
| | 10 ⁴ | 571.192 (0.039) | 571.347 (0.042) | 571.669 (0.039) | 1334.607 (0.111) | 1334.839 (0.111) | 1335.097 (0.111) | 3653.824 (0.437) | 3654.278 (0.442) | 3655.450 (0.437) |
| | 10 ⁵ | 571.209 (0.009) | 571.365 (0.008) | 571.646 (0.009) | 1334.585 (0.044) | 1334.818 (0.043) | 1335.105 (0.044) | 3653.643 (0.128) | 3654.096 (0.132) | 3655.500 (0.128) |
| | 10 ⁶ | 571.203 | 571.359 | 571.610 | 1334.600 | 1334.835 | 1335.135 | 3653.54 | 3654.00 | 3655.50 |
| | | | | | | | | | | |
| VAR(2) | 10 ² | 573.879 (0.445) | 574.301 (0.508) | 574.341 (0.285) | 1341.945 (2.231) | 1342.639 (2.263) | 1342.216 (1.013) | 3669.634 (4.805) | 3672.060 (4.689) | 3663.593 (1.896) |
| | 10 ³ | 573.832 (0.143) | 574.262 (0.176) | 574.915 (0.070) | 1342.509 (0.380) | 1343.318 (0.384) | 1343.802 (0.185) | 3665.465 (1.287) | 3666.773 (1.228) | 3667.902 (0.311) |
| | 10 ⁴ | 573.779 (0.043) | 574.222 (0.036) | 574.907 (0.057) | 1342.683 (0.132) | 1343.490 (0.128) | 1343.846 (0.130) | 3664.728 (0.513) | 3666.125 (0.501) | 3667.701 (0.173) |
| | 10 ⁵ | 573.811 (0.022) | 574.253 (0.021) | 574.903 (0.075) | 1342.492 (0.054) | 1343.296 (0.051) | 1343.763 (0.103) | 3664.445 (0.158) | 3665.845 (0.153) | 3667.307 (0.512) |
| | 10 ⁶ | 573.799 | 574.242 | 574.923 | 1342.492 | 1343.294 | 1343.653 | 3664.46 | 3665.86 | 3667.22 |
| | | | | | | | | | | |

Notes: Draws refers to the number of posterior draws and the number of draws in the reduced Gibbs step. Across-chain standard deviations are reported within brackets.

Table 6: LOG-MARGINAL DATA DENSITY

| Model | Draws | Four-variate | | Five-variate | | Six-variate | | | | |
|--------|-----------------|--------------------|--------------------|--------------------|---------------------|---------------------|---------------------|----------------------|----------------------|---------------------|
| | | Chib | Method 1 | Method 2 | Chib | Method 1 | Method 2 | Chib | Method 1 | Method 2 |
| VAR(3) | 10 ² | 576.414 (0.380) | 577.212 (0.362) | 577.647 (0.576) | 1349.777 (1.839) | 1351.338 (1.657) | 1348.766 (2.540) | 3677.069 (7.478) | 3680.538 (6.365) | 3673.852 (2.382) |
| | 10 ³ | 576.356 (0.148) | 577.239 (0.136) | 578.146 (0.116) | 1349.269 (0.683) | 1350.835 (0.670) | 1351.448 (0.117) | 3673.980 (1.908) | 3676.470 (1.903) | 3676.672 (0.275) |
| | 10 ⁴ | 576.338 (0.057) | 577.204 (0.043) | 578.096 (0.062) | 1349.074 (0.194) | 1350.703 (0.220) | 1351.396 (0.122) | 3672.860 (0.532) | 3675.708 (0.512) | 3677.231 (0.275) |
| | 10 ⁵ | 576.337 (0.018) | 577.201 (0.019) | 578.087 (0.063) | 1349.140 (0.077) | 1350.760 (0.076) | 1351.297 (0.097) | 3672.448 (0.281) | 3675.284 (0.275) | 3677.085 (0.244) |
| | 10 ⁶ | 576.341 | 577.207 | 578.125 | 1349.143 | 1350.770 | 1351.147 | 3672.57 | 3675.39 | 3676.81 |
| | | | | | | | | | | |
| VAR(4) | 10 ² | 576.581 (0.345) | 577.992 (0.437) | 578.536 (0.406) | 1353.412 (2.916) | 1355.789 (2.829) | 1353.021 (2.916) | 3683.978 (10.852) | 3690.937 (11.542) | 3671.483 (4.981) |
| | 10 ³ | 576.584 (0.127) | 578.028 (0.166) | 579.101 (0.087) | 1351.577 (0.489) | 1354.180 (0.525) | 1355.008 (0.212) | 3672.672 (1.743) | 3677.039 (1.793) | 3677.630 (0.870) |
| | 10 ⁴ | 576.524 (0.057) | 577.943 (0.040) | 578.998 (0.103) | 1351.576 (0.096) | 1354.259 (0.091) | 1355.032 (0.108) | 3671.367 (0.692) | 3675.923 (0.705) | 3678.001 (0.289) |
| | 10 ⁵ | 576.521 (0.015) | 77.950 (0.010) | 579.024 (0.041) | 1351.575 (0.039) | 1354.307 (0.035) | 1354.924 (0.100) | 3671.060 (0.180) | 3675.730 (0.209) | 3677.560 (0.155) |
| | 10 ⁶ | 576.513 | 577.946 | 579.058 | 1351.552 | 1354.286 | 1355.012 | 3670.99 | 3675.61 | 3677.55 |
| | | | | | | | | | | |

Notes: Draws refers to the number of posterior draws and the number of draws in the reduced Gibbs step. Across-chain standard deviations are reported within brackets.

Appendix

Appendix A provides the standard samplers used in the estimation. In Appendix B, we provide a detailed derivation of the Method 2. Appendix C shows that the posterior for the mean and the trend Γ of mean-adjusted VARs, discussed in Section 3.1, is conjugate to a Gaussian prior. In Appendix D we derive the analytical expression for conditional predictive density $p(Y|\Gamma)$, which is used to apply our estimators to mean-adjusted VARs of the form discussed in Section 3.1. Appendix E shows how to derive a close-form analytical expression for the conditional posteriors $\pi_{11}|(K, Y)$, $\pi_{22}|(K, Y)$ and $\Phi(j), \Sigma(j) | K, Y, j \in \{1, 2\}$ for Markov-Switching VARs in equation (29). In Appendix F, we prove that, conditional on factors, the posterior density for the parameter blocks in the factor model (40) equals their prior. This very last result has been used to show that the *analytical tractability condition* holds for Dynamic Factor Models in Section 37.

A Posterior samplers

Algorithm 1: Gibbs Sampler

Given an initial set of parameter values, $\Theta^{(0)}$, set $s = 0$ and perform the following steps

1. Draw $D^{(s+1)}$ from the conditional predictive density, $p(D|\Theta^{(s)}, Y)$
2. Draw $\theta_1^{(s+1)}$ from the conditional posterior, $p(\theta_1|\Theta_{>1}^{(s)}, D^{(s+1)}, Y)$
3. Draw $\theta_2^{(s+1)}$ from the conditional posterior, $p(\theta_2|\theta_1^{(s+1)}, \Theta_{>2}^{(s)}, D^{(s+1)}, Y)$
4. ...
5. Draw $\theta_m^{(s+1)}$ from the conditional posterior, $p(\theta_m|\theta_1^{(s+1)}, \dots, \theta_{m-1}^{(s+1)}, D^{(s+1)}, Y)$
6. Set $s = s + 1$. If $s \leq n_r$, go to step 1. Otherwise STOP.

Algorithm 2: Reduced-Gibbs Sampler:

Given an initial set of parameter values, $\Theta_{<i}^{(0)}$. Set $s = 0$ and perform the following steps

1. Draw $D^{(s+1)}$ from the conditional predictive density, $p(D|\Theta_{\leq i}^{(s)}, \tilde{\Theta}_{>i}, Y)$

2. If $i = 1$, then go to step 6. Else, draw $\theta_1^{(s+1)}$ from $p\left(\theta_1 | \Theta_{1 < \theta \leq i}^{(s)}, \tilde{\Theta}_{> i}, D^{(s+1)}, Y\right)$, where $\Theta_{1 < \theta \leq i}^{(s)} \equiv \{\theta_1, \dots, \theta_i\}$.
3. If $i = 2$, then go to step 6. Else, draw $\theta_2^{(s+1)}$ from $p\left(\theta_2 | \theta_1^{(s+1)} \Theta_{2 < \theta \leq i}^{(s)}, \tilde{\Theta}_{> i}, D^{(s+1)}, Y\right)$.
4. ...
5. If $i = (m - 1)$, then go to step 6. Else, draw $\theta_i^{(s+1)}$ from $p\left(\theta_i | \Theta_{1 \leq \theta \leq m-1}^{(s+1)}, \tilde{\Theta}_{> i}, D^{(s+1)}, Y\right)$.
6. Set $s = s + 1$. If $s \leq n_r$, go to step 1. Otherwise STOP.

Note that when $i = m$ the reduced Gibbs sampler coincides with the Gibbs sampler described in Algorithm 1.

B Derivation of Method 2

Gelfand and Dey (1994) propose the Reciprocal Importance Sampling (RIS) estimator to compute the MDD.

$$\hat{p}_{RIS}(Y) = \left[\frac{1}{N} \sum_{s=1}^N \frac{f(\theta^{(s)})}{k(\theta^{(s)}|Y)} \right]^{-1} \quad (46)$$

where $\{\theta^{(s)}\}_{s=1}^N$ are the posterior draws from the Gibbs sampler and $f(\cdot)$ stands for a weighting function, such that $\int f(\theta) d\theta = 1$. The RIS estimator is obtained as follows:

$$\frac{1}{p(Y)} = \int \frac{f(\theta)}{p(Y)} d\theta = \int \frac{f(\theta)}{k(\theta|Y)} p(\theta|Y) d\theta = \mathbb{E}_{p(\theta|Y)} \left[\frac{f(\theta)}{k(\theta|Y)} \right] \quad (47)$$

where $p(\theta|Y)$ is the posterior density and the second equality stems from the fact that the the MDD, using Bayes Theorem, can be expressed as $p(Y) = k(\theta|Y) / p(\theta|Y)$. A weighting function, $f(\cdot)$, that closely mimics the posterior kernel with thinner tails is desirable for the efficiency and the accuracy of the RIS estimator. Several weighting functions have been proposed in the literature. For example, Newton and Raftery (1999) use the prior as the weighting function. Gelfand and Dey (1994) propose a multivariate Student-t or Gaussian density with first and second moments estimated from the sample of posterior draws. Geweke (1999) suggests to use a truncated multivariate Gaussian density. We follow Geweke (1999) to construct one of our estimators. See Section 2.2.2.

We can exploit the result in (47) to write

$$\frac{1}{p(Y)} = \int f(\Theta_{>\tau}) \left(\frac{p(\Theta|Y)}{p(Y|\Theta)p(\Theta)} \right) d\Theta_{>\tau}$$

where the weighting function $f(\Theta_{>\tau})$ is known and such that $\int f(\Theta_{>\tau}) d\Theta_{>\tau} = 1$. We can rewrite the ratio within the bracket as follows:

$$\frac{1}{p(Y)} = \int f(\Theta_{>\tau}) \left(\frac{p(\Theta_{\leq\tau}|\Theta_{>\tau}, Y)}{p(Y|\Theta_{\leq\tau}, \Theta_{>\tau}) p(\Theta_{\leq\tau}|\Theta_{>\tau})} \frac{p(\Theta_{>\tau}|Y)}{p(\Theta_{>\tau})} \right) d\Theta_{>\tau} \quad (48)$$

The analytical expression of the conditional posterior density $\Theta_{\leq\tau} | (\Theta_{>\tau}, Y)$ is not available. The *analytical tractability condition*, however, ensures that we know the analytical form of the density $(\Theta_{\leq\tau}) | (\Theta_{>\tau}, D, Y)$. Note that

$$p(\Theta_{\leq\tau}|\Theta_{>\tau}, Y) = \int p(\Theta_{\leq\tau}|\Theta_{>\tau}, D, Y) p(D|\Theta_{>\tau}, Y) dD \quad (49)$$

By substituting the result in (49) into (48) we obtain

$$\frac{1}{p(Y)} = \int f(\Theta_{>\tau}) \left(\frac{\int p(\Theta_{\leq\tau}|\Theta_{>\tau}, D, Y) p(D|\Theta_{>\tau}, Y) dD p(\Theta_{>\tau}|Y)}{p(Y|\Theta_{\leq\tau}, \Theta_{>\tau}) p(\Theta_{\leq\tau}|\Theta_{>\tau}) p(\Theta_{>\tau})} \right) d\Theta_{>\tau}$$

One can observe that the densities outside the inner integral do not depend on D . Hence, we can re-write the equation above as

$$\frac{1}{p(Y)} = \int_D \int_{\Theta_{>\tau}} f(\Theta_{>\tau}) \left(\frac{p(\Theta_{\leq\tau} | (\Theta_{>\tau}, D, Y)) p(D|\Theta_{>\tau}, Y) p(\Theta_{>\tau}|Y)}{p(Y|\Theta_{\leq\tau}, \Theta_{>\tau}) p(\Theta_{\leq\tau}|\Theta_{>\tau}) p(\Theta_{>\tau})} \right) d(D, \Theta_{>\tau}) \quad (50)$$

where we have also reversed the order of integration.

Note that $p(D|\Theta_{>\tau}, Y) p(\Theta_{>\tau}|Y) = p(D, \Theta_{>\tau}|Y)$. Thus, we can write the equation in the main text.

$$\frac{1}{p(Y)} = \int f(\Theta_{>\tau}) \frac{p(\Theta_{\leq\tau}|\Theta_{>\tau}, D, Y)}{p(Y|\Theta_{\leq\tau}, \Theta_{>\tau}) p(\Theta_{\leq\tau}|\Theta_{>\tau}) p(\Theta_{>\tau})} p(D, \Theta_{>\tau}|Y) d(D, \Theta_{>\tau})$$

By using this result we can write:

$$\frac{1}{p(Y)} = \mathbb{E}_{p(D, \Theta_{>\tau}|Y)} \left[f(\Theta_{>\tau}) \frac{p(\Theta_{\leq\tau}|\Theta_{>\tau}, D, Y)}{p(Y|\Theta_{\leq\tau}, \Theta_{>\tau}) p(\Theta_{\leq\tau}|\Theta_{>\tau}) p(\Theta_{>\tau})} \right]$$

The *analytical tractability condition* implies that the analytical expression of the conditional predictive density, $p(\Theta_{\leq\tau}|\Theta_{>\tau}, D, Y)$, is known. Hence, we can estimate the marginal data density, $p(Y)$, through Method 2 as follows:

$$\hat{p}_{M2}(Y) = \left[\frac{1}{n} \sum_{s=1}^n \frac{p(\tilde{\Theta}_{\leq\tau}|\Theta_{>\tau}^{(s)}, D^{(s)}, Y)}{p(Y|\tilde{\Theta}_{\leq\tau}, \Theta_{>\tau}^{(s)}) p(\tilde{\Theta}_{\leq\tau}|\Theta_{>\tau}^{(s)}) p(\Theta_{>\tau}^{(s)})} f(\Theta_{>\tau}^{(s)}) \right]^{-1}$$

where the draws $\{\Theta_{>\tau}^{(s)}, D^{(s)}\}$ are the draws from the Gibbs sampler simulator (Algorithm 1) and $\tilde{\Theta}_{\leq\tau}$ is the posterior mode.

C Conditional Posterior $p(\Gamma|\Phi, \Sigma, Y)$

Note that

$$\begin{aligned} y_t &= \Gamma_1 + \Gamma_2 t + \tilde{y}_t \\ y_t - \sum_{j=1}^p \{\Phi\}_j \tilde{y}_{t-j} &= \Gamma_1 + \Gamma_2 t + \varepsilon_t \\ y_t - \sum_{j=1}^p \{\Phi\}_j (y_{t-j} - \Gamma_1 - \Gamma_2(t-j)) &= \Gamma_1 + \Gamma_2 t + \varepsilon_t \\ y_t - \sum_{j=1}^p \{\Phi\}_j y_{t-j} &= \left(I - \sum_{j=1}^p \{\Phi\}_j \right) \Gamma_1 \\ &\quad + \left(I \cdot t - \sum_{j=1}^p \{\Phi\}_j (t-j) \right) \Gamma_2 + \varepsilon_t \\ \check{y}(\Phi) &= A_t \Gamma + \varepsilon_t \end{aligned}$$

where $\check{y} \equiv y_t - \sum_{j=1}^p \{\Phi\}_j y_{t-j}$, $A_t \equiv \left[\left(I - \sum_{j=1}^p \{\Phi\}_j \right), \left(I \cdot t - \sum_{j=1}^p \{\Phi\}_j (t-j) \right) \right]$, $\Gamma \equiv \text{vec}(\Gamma')$.

Thus the kernel of the likelihood function is given by

$$-\frac{1}{2} \sum_{t=1}^T (\check{y} - A_t \Gamma)' \Sigma^{-1} (\check{y} - A_t \Gamma)$$

$$\begin{aligned}
&= -\frac{1}{2} \left[\sum_{t=1}^T \check{y}' \Sigma^{-1} \check{y} - 2 \left(\sum_{t=1}^T \check{y}' \Sigma^{-1} A_t \right) \Gamma + \Gamma' \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right) \Gamma \right] \\
&= -\frac{1}{2} \left[\Gamma - \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right]' \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right) \\
&\quad \left[\Gamma - \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right] \\
&\quad - \left(\sum_{t=1}^T \check{y}' \Sigma^{-1} A_t \right) \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) + \sum_{t=1}^T \check{y}' \Sigma^{-1} \check{y} \\
&= -\frac{1}{2} \left[\Gamma - \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right]' \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right) \\
&\quad \left[\Gamma - \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right] + \text{stuff not depending on } \theta_3
\end{aligned}$$

This suffices to conclude that the likelihood is Gaussian. We have to combine this likelihood with the prior $p(\Gamma)$, which is set to be:

$$p(\Gamma) \propto |V_\Gamma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\Gamma - \mu_\Gamma)' (V_\Gamma)^{-1} (\Gamma - \mu_\Gamma) \right\} \quad (51)$$

Hence the kernel of the posterior $p(\Gamma | \Phi, \Sigma, Y^T)$ will be

$$\begin{aligned}
&= -\frac{1}{2} \left\{ \left[\Gamma - \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right]' \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right) \right. \\
&\quad \cdot \left. \left[\Gamma - \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right] + (\Gamma - \mu_\Gamma)' (V_\Gamma)^{-1} (\Gamma - \mu_\Gamma) \right\}
\end{aligned}$$

$$\begin{aligned}
&= -\frac{1}{2}\left\{\Gamma' \left((V_\Gamma)^{-1} + \sum_{t=1}^T A_t' \Sigma^{-1} A_t \right) \Gamma - 2 \left[(\mu_\Gamma)' (V_\Gamma)^{-1} + \left(\sum_{t=1}^T \check{y}' \Sigma^{-1} A_t \right) \right] \Gamma \right. \\
&\quad \left. + \left(\sum_{t=1}^T \check{y}' \Sigma^{-1} A_t \right) \left(\sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) + (\mu_\Gamma)' (V_\Gamma)^{-1} \mu_\Gamma \right\} \\
&= -\frac{1}{2}\left\{\Gamma' \left((V_\Gamma)^{-1} + \sum_{t=1}^T A_t' \Sigma^{-1} A_t \right) \Gamma - 2 \left[(\mu_\Gamma)' (V_\Gamma)^{-1} + \left(\sum_{t=1}^T \check{y}' \Sigma^{-1} A_t \right) \right] \Gamma \right. \\
&\quad \left. + \text{stuff not depending on } \Gamma \right\} \\
&= -\frac{1}{2} \left[\Gamma - \left((V_\Gamma)^{-1} + \sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left[(V_\Gamma)^{-1} (\mu_\Gamma) + \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right] \right]' \\
&\quad \cdot \left((V_\Gamma)^{-1} + \sum_{t=1}^T A_t' \Sigma^{-1} A_t \right) \cdot \\
&\quad \cdot \left[\Gamma - \left((V_\Gamma)^{-1} + \sum_{t=1}^T A_t' \Sigma^{-1} A_t \right)^{-1} \left[(V_\Gamma)^{-1} (\mu_\Gamma) + \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right] \right]
\end{aligned}$$

Thus we can write the posterior of Γ as

$$\Gamma | \mathbf{Y}, \Phi, \Sigma \sim \mathcal{N}(\mu_\Gamma^T, V_\Gamma^T) \quad (52)$$

where

$$\begin{aligned}
V_\Gamma^T &= \left[(V_\Gamma)^{-1} + \sum_{t=1}^T A_t' \Sigma^{-1} A_t \right]^{-1} \\
\mu_\Gamma^T &= V_\Gamma^T \left[(V_\Gamma)^{-1} (\mu_\Gamma) + \left(\sum_{t=1}^T A_t' \Sigma^{-1} \check{y} \right) \right]
\end{aligned}$$

D Deriving the Conditional Predictive Density $p(Y|\Gamma)$

In this section, we derive the conditional predictive density $p(Y|\Gamma)$. Let us consider the VAR in deviation (13):

$$\tilde{\mathbf{Y}} = \Phi \mathbf{X} + \varepsilon$$

where $\tilde{\mathbf{Y}} \equiv [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_T]'$, $\mathbf{X} \equiv [x_1, x_2, \dots, x_T]'$, $x_t \equiv [\tilde{y}_{t-1}, \dots, \tilde{y}_{t-p}]$, and $\varepsilon \equiv [\varepsilon_1, \dots, \varepsilon_T]'$.

Since this prior is constructed via dummy observations. Then the marginal data density can be expressed as

$$p(\tilde{\mathbf{Y}}|\tilde{\mathbf{Y}}^*) = \frac{\int P(\tilde{\mathbf{Y}}, \tilde{\mathbf{Y}}^*|\Phi, \Sigma) p(\Phi, \Sigma) d\Sigma d\Phi}{\int P(\tilde{\mathbf{Y}}^*|\Phi, \Sigma) p(\Phi, \Sigma) d\Sigma d\Phi} \quad (53)$$

Then we need to compute

$$\int P(\tilde{\tilde{\mathbf{Y}}|\Phi, \Sigma) p(\Phi, \Sigma) d\Sigma d\Phi \quad (54)$$

where the vector $\tilde{\tilde{\mathbf{Y}}} = (\tilde{\mathbf{Y}}, \tilde{\mathbf{Y}}^*)$.

The likelihood can be expressed as

$$P(\tilde{\tilde{\mathbf{Y}}|\Phi, \Sigma) = (2\pi)^{-\frac{Tn}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\} \quad (55)$$

$$\times \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\Gamma - \hat{\Phi}\right)' \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \left(\Phi - \hat{\Phi}\right)\right]\right\} \quad (56)$$

where T is the number of rows of $\tilde{\tilde{\mathbf{Y}}}$ and n is the number of columns of $\tilde{\tilde{\mathbf{Y}}}$ and

$$\hat{\Phi} = \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}' \tilde{\tilde{\mathbf{Y}}} \quad (57)$$

$$\mathbf{S} = \left(\tilde{\tilde{\mathbf{Y}}} - \tilde{\mathbf{X}} \hat{\Phi}\right)' \left(\tilde{\tilde{\mathbf{Y}}} - \tilde{\mathbf{X}} \hat{\Phi}\right) \quad (58)$$

The prior we used is an improper one as

$$p(\Phi, \Sigma) = |\Sigma|^{-\frac{n+1}{2}} \quad (59)$$

Combining equation (55) and equation (59) yields

$$P(\tilde{\tilde{\mathbf{Y}}|\Phi, \Sigma) p(\Phi, \Sigma) = (2\pi)^{-\frac{Tn}{2}} |\Sigma|^{-\frac{T+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\} \quad (60)$$

$$\times \exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\Phi - \hat{\Phi}\right)' \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \left(\Phi - \hat{\Phi}\right)\right]\right\}$$

So if we now we integrate the equation (60) across Σ and Φ we get the marginal data density as indicated in equation (54):

$$\int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi \quad (61)$$

$$= \int (2\pi)^{-\frac{Tn}{2}} |\Sigma|^{-\frac{T+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\} \quad (62)$$

$$\exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\Phi - \hat{\Phi}\right)' \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \left(\Phi - \hat{\Phi}\right)\right]\right\} d\Sigma d\Phi \quad (63)$$

By multiplying and dividing by $\left|\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right|^{\frac{1}{2}}$ inside the integral to get

$$\int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi \quad (64)$$

$$= (2\pi)^{-\frac{Tn}{2}} \int |\Sigma|^{-\frac{n+1}{2}} |\Sigma|^{-\frac{T}{2}} \left|\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\} \cdot \left|\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right|^{-\frac{1}{2}} \quad (65)$$

$$\exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\Phi - \hat{\Phi}\right)' \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \left(\Phi - \hat{\Phi}\right)\right]\right\} d\Sigma d\Phi \quad (66)$$

Note that it can be show that:

$$tr\left[\Sigma^{-1}\left(\Phi - \hat{\Phi}\right)' \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \left(\Phi - \hat{\Phi}\right)\right] = (\varphi_2 - \hat{\varphi}_2)' \left[\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right]^{-1} (\varphi_2 - \hat{\varphi}_2) \quad (67)$$

where $\varphi \equiv vec(\Phi)$ and $\hat{\varphi}_2 \equiv vec(\hat{\Phi})$. Hence we can write

$$\int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi \quad (68)$$

$$= (2\pi)^{-\frac{Tn}{2}} \int |\Sigma|^{-\frac{n+1}{2}} |\Sigma|^{-\frac{T}{2}} \left|\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\} \cdot \left|\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right|^{-\frac{1}{2}} \quad (69)$$

$$\exp\left\{-\frac{1}{2}(\varphi_2 - \hat{\varphi}_2)' \left[\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right]^{-1} (\varphi_2 - \hat{\varphi}_2)\right\} d\Sigma d\Phi \quad (70)$$

Recall that

$$\left|\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right|^{\frac{1}{2}} = |\Sigma|^{\frac{k}{2}} \left|\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right|^{-\frac{n}{2}} \quad (71)$$

where k is equal to the number of columns of $\tilde{\mathbf{X}}$. Thus, if there is a *constant* in the VAR, then $k = np + 1$, Otherwise, $k = np$. It follows that:

$$\int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi \quad (72)$$

$$= (2\pi)^{-\frac{Tn}{2}} \int |\Sigma|^{-\frac{n+1}{2}} |\Sigma|^{-\frac{T}{2}} |\Sigma|^{\frac{k}{2}} \left|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\} \cdot \left|\Sigma \otimes \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right|^{-\frac{1}{2}} \quad (73)$$

$$\exp\left\{-\frac{1}{2}(\varphi_2 - \hat{\varphi}_2)' \left[\Sigma \otimes \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1} (\varphi_2 - \hat{\varphi}_2)\right\} d\Sigma d\Phi \quad (74)$$

$$= (2\pi)^{-\frac{Tn}{2}} \left|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right|^{-\frac{n}{2}} \int |\Sigma|^{-\frac{(T-k)+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\} \cdot \left|\Sigma \otimes \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right|^{-\frac{1}{2}} \quad (75)$$

$$\exp\left\{-\frac{1}{2}(\varphi_2 - \hat{\varphi}_2)' \left[\Sigma \otimes \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1} (\varphi_2 - \hat{\varphi}_2)\right\} d\Sigma d\Phi \quad (76)$$

If we define $v \equiv T - k$ and then multiply and divide by $(2^{\frac{vn}{2}} \Gamma_n(\frac{v}{2})) |\mathbf{S}|^{\frac{v}{2}}$, we get

$$\int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi \quad (77)$$

$$= (2\pi)^{-\frac{Tn}{2}} \left|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right|^{-\frac{n}{2}} |\mathbf{S}|^{-\frac{v}{2}} \cdot 2^{\frac{vn}{2}} \Gamma_n\left(\frac{v}{2}\right) \int \frac{|\mathbf{S}|^{\frac{v}{2}} |\Sigma|^{-\frac{v+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\}}{2^{\frac{vn}{2}} \Gamma_n\left(\frac{v}{2}\right)} \cdot \left|\Sigma \otimes \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right|^{-\frac{1}{2}} \quad (78)$$

$$\exp\left\{-\frac{1}{2}(\varphi_2 - \hat{\varphi}_2)' \left[\Sigma \otimes \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1} (\varphi_2 - \hat{\varphi}_2)\right\} d\Sigma d\Phi \quad (79)$$

where $\Gamma_n\left(\frac{v}{2}\right)$ is the multivariate gamma function:

$$\Gamma_n\left(\frac{v}{2}\right) = \pi^{\frac{n(n-1)}{4}} \prod_{j=1}^n \Gamma\left(\frac{v}{2} + \frac{(1-j)}{2}\right) \quad (80)$$

Finally, multiply and divide inside the integral operator by $(2\pi)^{-\frac{kn}{2}}$, we get

$$\int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi \quad (81)$$

$$= (2\pi)^{-\frac{(T-k)n}{2}} \left| \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \right|^{-\frac{n}{2}} |\mathbf{S}|^{-\frac{v}{2}} \cdot 2^{\frac{vn}{2}} \Gamma_n\left(\frac{v}{2}\right) \quad (82)$$

$$\int \frac{|\mathbf{S}|^{\frac{v}{2}} |\Sigma|^{-\frac{v+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\}}{2^{\frac{vn}{2}} \Gamma_n\left(\frac{v}{2}\right)} \cdot (2\pi)^{-\frac{nk}{2}} \left| \Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1} \right|^{-\frac{1}{2}} \quad (83)$$

$$\exp\left\{-\frac{1}{2}(\varphi_2 - \hat{\varphi}_2)' \left[\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right]^{-1} (\varphi_2 - \hat{\varphi}_2)\right\} d\Sigma d\Phi \quad (84)$$

Note that $|\mathbf{S}|^{\frac{v}{2}} |\Sigma|^{-\frac{v+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma\mathbf{S}]\right\}$ does not depend on Φ . Hence, we can put it outside the integral operator taken with respect to Φ and actually get

$$\begin{aligned} & \int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi \\ = & (2\pi)^{-\frac{(T-k)n}{2}} \left| \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \right|^{-\frac{n}{2}} |\mathbf{S}|^{-\frac{v}{2}} 2^{\frac{vn}{2}} \cdot \Gamma_n\left(\frac{v}{2}\right) \int \frac{|\mathbf{S}|^{\frac{v}{2}} |\Sigma|^{-\frac{v+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\}}{2^{\frac{vn}{2}} \cdot \Gamma_n\left(\frac{v}{2}\right)} \cdot \\ & \left\{ \int (2\pi)^{-\frac{nk}{2}} \left| \Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1} \right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\varphi_2 - \hat{\varphi}_2)' \left[\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right]^{-1} (\varphi_2 - \hat{\varphi}_2)\right\} d\Phi \right\} \end{aligned}$$

Since the expression inside the inner integral is a normal we get the following

$$\int (2\pi)^{-\frac{nk}{2}} \left| \Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1} \right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\varphi_2 - \hat{\varphi}_2)' \left[\Sigma \otimes \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}}\right)^{-1}\right]^{-1} (\varphi_2 - \hat{\varphi}_2)\right\} d\Phi = 1$$

So we get

$$\begin{aligned} & \int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi \\ = & (2\pi)^{-\frac{(T-k)n}{2}} \left| \tilde{\mathbf{X}}' \tilde{\mathbf{X}} \right|^{-\frac{n}{2}} |\mathbf{S}|^{-\frac{v}{2}} 2^{\frac{vn}{2}} \cdot \Gamma_n\left(\frac{v}{2}\right) \int \frac{|\mathbf{S}|^{\frac{v}{2}} |\Sigma|^{-\frac{v+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma\mathbf{S}]\right\}}{2^{\frac{vn}{2}} \cdot \Gamma_n\left(\frac{v}{2}\right)} d\Sigma \end{aligned}$$

Since the argument of the integral is an inverted-wishart distribution, we have that

$$\int \frac{|\mathbf{S}|^{\frac{v}{2}} |\Sigma|^{-\frac{v+n+1}{2}} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\mathbf{S}]\right\}}{2^{\frac{vn}{2}} \cdot \Gamma_n\left(\frac{v}{2}\right)} d\Sigma = 1 \quad (85)$$

and then

$$\int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi = (2\pi)^{-\frac{(T-k)n}{2}} \left|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right|^{-\frac{n}{2}} |\mathbf{S}|^{-\frac{v}{2}} 2^{\frac{vn}{2}} \cdot \pi^{\frac{n(n-1)}{4}} \prod_{j=1}^n \Gamma\left(\frac{v}{2} + \frac{(1-j)}{2}\right) \quad (86)$$

and then by noticing that $2^{-\frac{(T-k)n}{2}} 2^{\frac{vn}{2}}$ cancels out as $v = T - k$ we, finally, get

$$\int P\left(\tilde{\mathbf{Y}}|\Phi, \Sigma\right) p(\Phi, \Sigma) d\Sigma d\Phi = \pi^{-\frac{(T-k)n}{2}} \left|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right|^{-\frac{n}{2}} |\mathbf{S}|^{-\frac{v}{2}} \cdot \pi^{\frac{n(n-1)}{4}} \prod_{j=1}^n \Gamma\left(\frac{v}{2} + \frac{(1-j)}{2}\right) \quad (87)$$

where

$$\pi^{\frac{n(n-1)}{4}} \prod_{j=1}^n \Gamma\left(\frac{v}{2} + \frac{(1-j)}{2}\right) = \Gamma_n\left(\frac{v}{2}\right)$$

Equation (87) is the analytical expression for the conditional predictive density, $p(\mathbf{Y}|\Gamma)$.

E MS VARs

Let t_{jj} be the number of periods the system is in regime j and $t_{ij}, i \neq j$ the number of times the system switches from state i to state j . Let us suppose that the history of regimes K implies k breaks that happen at $t \in \{T_1, \dots, T_k\}$. Without loss of generality, let us assume that $K_1 = 1$ and $K_T = 1$.³⁰ Furthermore, we define the following matrices:

$$\begin{aligned} Y(1) &= [y_1, \dots, y_{T_1}, y_{T_2+1}, \dots, y_{T_3}, y_{T_4+1}, \dots, y_{T_{k-1}+1}, \dots, y_T]' \\ X(1) &= [x_1, \dots, x_{T_1}, x_{T_2+1}, \dots, x_{T_3}, x_{T_4+1}, \dots, x_{T_{k-1}+1}, \dots, x_T]' \\ Y(2) &= [y_{T_1+1}, \dots, y_{T_2}, y_{T_3+1}, \dots, y_{T_4}, y_{T_5+1}, \dots, y_{T_{k-2}+1}, \dots, y_{T_{k-1}}]' \\ X(2) &= [x_{T_1+1}, \dots, x_{T_2}, x_{T_3+1}, \dots, x_{T_4}, x_{T_5+1}, \dots, x_{T_{k-2}+1}, \dots, x_{T_{k-1}}]' \end{aligned}$$

with $x_t = [1, y'_{t-1}, \dots, y'_{t-p}]'$. Let $T_0(j)$ be the number of dummy observations, $T(j)$ be the total number of periods the system is in regime j in the history K , and $\bar{Y}(j) = [Y^*(j)', Y(j)']'$, $\bar{X} = [X^*(j)', X(j)']'$.

Given that $\tau = m$, equation (3), which characterizes Method 1, reduces to

$$\hat{p}_{M1}(Y) = \frac{p(Y|\tilde{\Theta}) p(\tilde{\Theta})}{p(\tilde{\Theta}|Y)} \quad (88)$$

³⁰Having that $K_T = 2$, for instance, would only cause the definition of matrices $Y(j)$ and $X(j)$, $j \in \{1, 2\}$ to change in a straightforward manner.

where $\tilde{\Theta} \equiv \left[(\tilde{\pi}_{jj})_{j \in \{1,2\}}, (\tilde{\Phi}(j), \tilde{\Sigma}(j))_{j \in \{1,2\}} \right]$ is the posterior mode. Method 1 approximates the joint posterior density $p(\tilde{\Theta}|Y)$ from the output of the Gibbs sampler as follows:

$$p(\tilde{\Theta}|Y) = \frac{1}{n_r} \sum_{s=1}^{n_r} \prod_{j=1}^2 p(\tilde{\pi}_{jj}|K^{(s)}, Y) \cdot p(\tilde{\Phi}(j), \tilde{\Sigma}(j)|K^{(s)}, Y) \quad (89)$$

where $p(\pi_{11}|K^{(s)}, Y) \sim \text{Beta}(t_{11}^* + t_{11}, t_{12}^* + t_{12})$, $p(\pi_{22}|K^{(s)}, Y) \sim \text{Beta}(t_{22}^* + t_{22}, t_{21}^* + t_{21})$ and $p(\Phi(j), \Sigma(j)|K^{(s)}, Y) \sim \text{MNIW}(\hat{\Phi}(j), (\bar{X}(j)' \bar{X}(j))^{-1}, \hat{S}(j), \bar{T}(j) - k)$, with $\bar{T}(j) = T_0(j) + T(j)$, $k = np + 1$, $\hat{\Phi}(j) = (\bar{X}(j)' \bar{X}(j))^{-1} \bar{X}(j)' \bar{Y}(j)$, and $\hat{S}(j) = (\bar{Y}(j) - \bar{X}(j) \hat{\Phi}(j))' (\bar{Y}(j) - \bar{X}(j) \hat{\Phi}(j))$.

F Posterior for the Parameter of the Factor Model

We want to show that conditional to the factors the posterior density for the parameter blocks in the factor model equals their prior. The joint posterior for the four parameter blocks of the DFM model (39)-(40) is given by

$$p(\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1|F, Y) = \frac{p(Y|\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1, F) \cdot p(\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1|F)}{\int p(Y|\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1, F) \cdot p(\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1|F) d(\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1)}$$

Note that conditional on the factors F , the likelihood $p(Y|\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1, F)$ simplifies to $p(Y|\Phi_1, \Sigma_1, F)$. Hence,

$$p(\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1|F, Y) = \frac{p(Y|\Phi_1, \Sigma_1, F) \cdot p(\Phi_1, \Sigma_1)}{\int p(Y|\Phi_1, \Sigma_1, F) \cdot p(\Phi_1, \Sigma_1) d(\Phi_1, \Sigma_1)} \cdot \frac{p(\Phi_0, \varepsilon_0)}{\int p(\Phi_0, \varepsilon_0) d(\Phi_0, \varepsilon_0)}$$

where also we use the assumption made on prior specification made in Section 3.5. Furthermore, note that $\int p(\Phi_0, \varepsilon_0) d(\Phi_0, \varepsilon_0) = 1$. It follows that

$$p(\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1|F, Y) = p(\Phi_1, \Sigma_1|F, Y) \cdot p(\Phi_0, \varepsilon_0) \quad (90)$$

Recall that the posterior for the parameters in the factor model in (40) is defined as

$$p(\Phi_0, \varepsilon_0|F, Y) = \int p(\Phi_0, \varepsilon_0, \Phi_1, \Sigma_1|F, Y) d(\Phi_1, \Sigma_1)$$

Using the result in equation (90), we obtain

$$p(\Phi_0, \varepsilon_0|F, Y) = \int p(\Phi_1, \Sigma_1|F, Y) \cdot p(\Phi_0, \varepsilon_0) d(\Phi_1, \Sigma_1)$$

Since $\int p(\Phi_1, \Sigma_1|F, Y) d(\Phi_1, \Sigma_1) = 1$, then $p(\Phi_0, \varepsilon_0|F, Y) = p(\Phi_0, \varepsilon_0)$, which is what we wanted to show.